A coalescent dual process for a Wright-Fisher diffusion with recombination and its application to haplotype partitioning

Robert C. Griffiths[a], Paul A. Jenkins[b,c,*], Sabin Lessard[d]

[a]*Department of Statistics, University of Oxford, United Kingdom*
[b]*Department of Statistics, University of Warwick, United Kingdom*
[c]*Department of Computer Science, University of Warwick, United Kingdom*
[d]*Département de Mathématiques et de Statistique, Université de Montréal, Montréal, Canada*

Abstract

Duality plays an important role in population genetics. It can relate results from forwards-in-time models of allele frequency evolution with those of backwards-in-time genealogical models; a well known example is the duality between the Wright-Fisher diffusion for genetic drift and its genealogical counterpart, the coalescent. There have been a number of articles extending this relationship to include other evolutionary processes such as mutation and selection, but little has been explored for models also incorporating crossover recombination. Here, we derive from first principles a new genealogical process which is dual to a Wright-Fisher diffusion model of drift, mutation, and recombination. The process is reminiscent of the *ancestral recombination graph*, a widely-used multilocus genealogical model, but here ancestral lineages are typed and transition rates are regarded as being conditioned on an observed configuration at the leaves of the genealogy. Our approach is based on expressing a putative duality relationship between two models via their infinitesimal generators, and then seeking an appropriate test function to ensure the validity of the duality equation. This approach is quite general, and we use it to find dualities for several important variants, including both a discrete L-locus model of a gene and a continuous model in which mutation and recombination events are scattered along the gene according to

*Corresponding author. Address: Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom. E-mail: p.jenkins@warwick.ac.uk

continuous distributions. As an application of our results, we derive a series expansion for the transition function of the diffusion. Finally, we study in further detail the case in which mutation is absent. Then the dual process describes the dispersal of ancestral genetic material across the ancestors of a sample. The stationary distribution of this process is of particular interest; we show how duality relates this distribution to haplotype fixation probabilities. We develop an efficient method for computing such probabilities in multilocus models.

Keywords: coalescent, Wright-Fisher diffusion, recombination, duality

---

1. Introduction

The concept of duality is a powerful technique for inferring the properties of one Markov process by looking at another related process, usually (as in this paper) discovered by considering the dynamics of the former in reverse time (see Jansen and Kurt, 2014, for recent review). The idea has found many applications in population genetics, playing for example a central role in the constructions of the ancestral selection graph (Krone and Neuhauser, 1997; Neuhauser and Krone, 1997) and the ancestral influence graph (Donnelly and Kurtz, 1999). One particularly well known duality is between the Wright-Fisher diffusion describing pure genetic drift and Kingman's coalescent (Kingman, 1982). To illustrate the idea, consider a single neutral locus with two alleles. The Wright-Fisher diffusion $(X_t)_{t\geq 0}$ is the process on $[0,1]$ describing the evolution of the frequency of one allele, with infinitesimal generator

$$\mathscr{L}f(x) = \frac{1}{2}x(1-x)f''(x) \tag{1}$$

and domain $\mathscr{D}(\mathscr{L}) = C^2([0,1])$. The corresponding dual is the pure death process $(L_t)_{t\geq 0}$ on $\mathbb{N} = \{0, 1, \dots\}$ with infinitesimal generator

$$\mathscr{K}f(n) = \binom{n}{2}[f(n-1) - f(n)], \tag{2}$$

which describes the dynamics of the *ancestral*, or *block-counting*, process of Kingman's coalescent.

The two processes are dual with respect to the function $F : [0,1] \times \mathbb{N} \to \mathbb{R}$ defined by $F(x,n) = x^n$ (i.e. *moment duals*): for each $x \in [0,1]$, $n \in \mathbb{N}$ and

$t \geq 0$,

$$\mathbb{E}[F(X_t, n) \mid X_0 = x] = \mathbb{E}[F(x, L_t) \mid L_0 = n]. \tag{3}$$

We note for later use that this implies

$$\mathscr{L} F(\cdot, n)(x) = \mathscr{K} F(x, \cdot)(n), \qquad x \in [0,1], \quad n \in \mathbb{N}, \tag{4}$$

and for general $\mathscr{L}$, $\mathscr{K}$, and F, the converse is also true under certain conditions on F (Jansen and Kurt, 2014). We also emphasise that, in this example and all others encountered in this paper, this duality is obtained via time-reversal, so that the time indices in the two processes run in different directions. Were we to run the two processes on a joint probability space, running X_t from time 0 to T would correspond to running L_t *backwards* from time T to 0.

There have been numerous extensions to the models captured by (4). For example, Ethier and Griffiths (1990a) describe a birth-death process which is dual to a two-locus Wright-Fisher diffusion with recombination between the two loci, and use it to prove an ergodic theorem for the diffusion. Mano (2013) uses the same process to derive a method to compute the transient moments of the diffusion. Generalising further, Ethier and Kurtz (1993) describe a duality relationship between a Fleming-Viot process with very general mutation, selection, and recombination operators and a function-valued dual process analogous to the block-counting process of the coalescent. Here, the function changes state as a jump process reminiscent of (2) due to genetic drift, selection, and recombination, while mutation contributes a deterministic component evolving the function continuously between jumps. Dualities in which mutation is either deterministic or absent can be used to compute some quantities of interest in the two models, but they are not the most general available. In this paper our purpose is different: it is to develop a coalescent dual for the Wright-Fisher diffusion in which mutation contributes to the *random* evolution of the dual process. This type of duality is important because the dual process describes the posterior genealogical dynamics of a sample, conditional on the allelic configuration observed in the present day. This is precisely the process of interest when one wishes to perform statistical inference under a coalescent model given some sample of genetic variation taken from a contemporary population (see Stephens, 2007, for an introduction). For example, a careful approximation of these dynamics provides a suitable proposal process in an importance sampling algorithm (examples for multilocus models include Griffiths and Marjoram,

1996; Fearnhead and Donnelly, 2001; Larribe et al., 2002; Griffiths et al., 2008; Larribe and Lessard, 2008; Jenkins and Griffiths, 2011; Kamm et al., 2016). This duality is also important because it provides a way of obtaining an expression for the transition function of the underlying diffusion (Griffiths, 1979; Donnelly and Tavaré, 1987; Ethier and Griffiths, 1993).

Dualities of this latter form have been developed for a number of models extending (1) and (2). These include models of mutation (Griffiths, 1980; Donnelly and Tavaré, 1987), natural selection (Barbour et al., 2000; Fearnhead, 2002; Stephens and Donnelly, 2003; Etheridge and Griffiths, 2009), and Λ-coalescent dynamics (Etheridge et al., 2010), as well as dualities for the Moran model which is a prelimit of the corresponding diffusion (Etheridge and Griffiths, 2009; Etheridge et al., 2010). Hitherto, there has not been described a corresponding dual process for models incorporating both mutation and recombination (by which we mean homologous, meiotic, crossover). [The existence of one such process is implicit in Fearnhead and Donnelly (2001) and Griffiths et al. (2008), but there the focus was on inference rather than any description of the process.] The goal of this paper is to derive such a duality relationship from first principles: in particular, we identify a genealogical dual for the Wright-Fisher diffusion with recombination which is similar to the ancestral recombination graph (ARG) of Griffiths and Marjoram (1997); the key differences being that here the lineages are typed, and jumps in the genealogical process are to be understood in an *a posteriori* sense. We obtain results both for a finite-locus model with general mutation structure and for its limit with continuous breakpoint distribution and infinitely-many-sites mutation. Our key object of study is a generalisation of the generator $\mathscr{L}$ defined in (1) and the duality identity (4). As applications of our approach we recover systems of recursive equations for the sampling distribution of the models (usually obtained more toilsomely by direct coalescent arguments), and we also obtain the first transition function expansion for a diffusion model incorporating recombination. Finally, we study the case of no mutation in further detail and develop an efficient method for computing the distribution of how ancestral genetic material is dispersed across the ancestors of a contemporary population (the so-called *partitioning process*). Using duality, these distributions also yield fixation probabilities for haplotypes in multilocus models.

The paper is structured as follows. In Section 2 we illustrate our approach with a known example of a K-allele system at a single locus. We then extend this in Section 3 to an L-locus model. In Section 4 we apply these results

to develop a series expansion for the transition function of the diffusion. In Section 5 we generalise the model further, to a continuous model of a gene in which mutation and recombination rates are modelled by a probability density function. In Section 6 we return to the L-locus model and study in further detail the dual process of a Wright-Fisher diffusion without mutation, and Section 7 concludes with a brief discussion.

2. Warm up: K-alleles at one locus

To illustrate the main idea and to clarify some notation, we first consider an extension of (4) to incorporate K-alleles with parent-independent mutation (PIM) at one locus. The key step is to make a judicious choice of duality function F so that, when we apply to it the infinitesimal generator of the underlying diffusion as an operator on the first variable of F, we *recognise* the resulting expression as the action of another generator acting on the second variable. Further applications of this idea can be found in Ethier and Griffiths (1993), Barbour et al. (2000), and Etheridge and Griffiths (2009).

Denote the finite type space of the locus by $E = \{1, \dots, K\} =: [K]$. The mutation model is specified by a rate parameter $\theta > 0$ and a distribution $(P_i)_{i\in E}$ over the type of a mutant offspring (independent of the parental allele). Within this framework, the Wright-Fisher diffusion $\boldsymbol{X} = (\boldsymbol{X}_t)_{t\geq 0}$ has state space

$$\Delta_E = \left\{ \boldsymbol{x} = (x_i)_{i\in E} \in [0,1]^E : \sum_{i\in E} x_i = 1 \right\} \tag{5}$$

and generator

$$\mathscr{L} f(\boldsymbol{x}) = \frac{1}{2} \sum_{i\in E} \sum_{j\in E} x_i(\delta_{ij} - x_j) \frac{\partial^2}{\partial x_i \partial x_j} f(\boldsymbol{x}) + \frac{\theta}{2} \sum_{i\in E} (P_i - x_i) \frac{\partial}{\partial x_i} f(\boldsymbol{x}), \tag{6}$$

where δ_{ij} denotes the Kronecker delta, and $\mathscr{D}(\mathscr{L}) = C^2(\Delta_E)$. Motivated by the choice of $F(x, n)$ we encountered above, let us evaluate $\mathscr{L} F(\boldsymbol{x}, \boldsymbol{n})$ for $F : \Delta_E \times \mathbb{N}^{|E|} \to \mathbb{R}$ defined by

$$F(\boldsymbol{x}, \boldsymbol{n}) = \frac{1}{m(\boldsymbol{n})} \prod_{i\in E} x_i^{n_i}, \tag{7}$$

for some $m : \mathbb{N}^{|E|} \to \mathbb{R}$ yet to be determined (here, $\boldsymbol{n} = (n_1, n_2, \ldots, n_K) \in \mathbb{N}^K$ and $|Z|$ denotes the cardinality of a set Z). We find

$$\mathscr{L}F(\cdot, \boldsymbol{n})(\boldsymbol{x}) = \\ \sum_{i \in E} \frac{n_i(n_i + \theta P_i - 1)}{2} \frac{m(\boldsymbol{n} - \boldsymbol{e}_i)}{m(\boldsymbol{n})} F(\boldsymbol{x}, \boldsymbol{n} - \boldsymbol{e}_i) - \frac{n(n + \theta - 1)}{2} F(\boldsymbol{x}, \boldsymbol{n}), \quad (8)$$

where $\boldsymbol{e}_i = (\delta_{ij})_{j=1,\ldots,K}$. This can be interpreted as the generator of a pure jump process evolving $\boldsymbol{n}$ on $\mathbb{N}^K$ if we can choose $m(\boldsymbol{n})$ so that (8) is in the form

$$\mathscr{L}F(\cdot, \boldsymbol{n})(\boldsymbol{x}) = \sum_{\widehat{\boldsymbol{n}}} q(\boldsymbol{n}, \widehat{\boldsymbol{n}})[F(\boldsymbol{x}, \widehat{\boldsymbol{n}}) - F(\boldsymbol{x}, \boldsymbol{n})], \quad (9)$$

where $\boldsymbol{Q} = (q(\cdot, \cdot))$ is a rate matrix; that is, it has negative diagonal elements, nonnegative off-diagonal elements, and rows summing to 0. Now, take expectations in (9) with respect to the stationary distribution of $\boldsymbol{X}$ and use the identity

$$\mathbb{E}[\mathscr{L}F(\boldsymbol{X}_\infty, \boldsymbol{n})] = 0 \quad (10)$$

to obtain

$$\boldsymbol{0} = \boldsymbol{Q}\boldsymbol{v}, \quad (11)$$

where $\boldsymbol{v} = (\mathbb{E}[F(\boldsymbol{X}_\infty, \cdot)])$ is a column vector. Here we generically use $\boldsymbol{X}_\infty$ to denote the process at stationarity. Equation (11) can be ensured if $\boldsymbol{v}$ has identical entries. In other words, we should choose $m(\boldsymbol{n})$ in (7) so that $\mathbb{E}[F(\boldsymbol{X}_\infty, \boldsymbol{n})]$ is a constant (and without loss of generality, 1). Using that $\boldsymbol{X}_\infty \sim \text{Dirichlet}(\theta P_1, \theta P_2, \ldots, \theta P_K)$ (Wright, 1949), we find by taking expectation in (7) that we require

$$m(\boldsymbol{n}) = \mathbb{E}\left[\prod_{i \in E} (\boldsymbol{X}_\infty)_i^{n_i}\right] = \frac{\prod_{i \in E} (\theta P_i)_{n_i}}{(\theta)_n}, \quad (12)$$

where, for $\phi \in \mathbb{R}_{\geq 0}$, $(\phi)_n := \phi(\phi+1)\ldots(\phi+n-1)$ denotes the nth ascending factorial of ϕ (and $(\phi)_0 := 1$). Then (8) becomes

$$\mathscr{L}F(\cdot, \boldsymbol{n})(\boldsymbol{x}) = \sum_{i \in E} \frac{n_i(n + \theta - 1)}{2} F(\boldsymbol{x}, \boldsymbol{n} - \boldsymbol{e}_i) - \frac{n(n + \theta - 1)}{2} F(\boldsymbol{x}, \boldsymbol{n}),$$

which is of the required form. (Perhaps surprisingly, this result suggests that the generator of the dual process does not depend on the P_i. However, the P_i do appear in the function F, which is not just an arbitrary function.)

In summary, the diffusion with generator (6) is dual to a pure death process on $\mathbb{N}^K$ with transition rate matrix

$$q(\boldsymbol{n}, \widehat{\boldsymbol{n}}) = \frac{n+\theta-1}{2} \times \begin{cases} n_i & \text{if } \widehat{\boldsymbol{n}} = \boldsymbol{n} - \boldsymbol{e_i}, \\ -n, & \text{if } \widehat{\boldsymbol{n}} = \boldsymbol{n}, \end{cases} \tag{13}$$

and the duality function is

$$F(\boldsymbol{x}, \boldsymbol{n}) = \frac{(\theta)_n}{\prod_{i \in E} (\theta P_i)_{n_i}} \prod_{i \in E} x_i^{n_i}. \tag{14}$$

From (13), an interpretation of the dual process is as follows: at rate $n(n+\theta-1)/2$, choose a gene to coalesce or mutate. At the chosen event, the gene involved is of type i with probability n_i/n. It is well known that, under a PIM model, the posterior probability that any particular lineage was involved in the most recent event is independent of its type. At either type of event, the lineage involved is lost, which is reminiscent of coalescent simulation under the *prior*: (only) under a PIM model, lineages undergoing mutation can be killed, so a simulated coalescent history becomes a random forest with each tree describing the genealogy of the sampled descendants of a single mutant.

Inspection of (14) might lead one to suspect that the duality between the two processes is really about equivalence of sampling distributions. Let us unpick this further by plugging (14) into the duality equation (3) and comparing the two sides. We contend that we have obtained two ways of addressing the following: What is the ratio of (i) the probability that a random sample of size n results in an ordered allelic configuration which, when unordered, yields the vector $\boldsymbol{n}$, given that the population allele frequencies a time t ago were $\boldsymbol{x}$; and (ii) the same probability without this extra information about the population? Using (12), the left of (3) is

$$\mathbb{E}[F(\boldsymbol{X}_t, \boldsymbol{n}) \mid \boldsymbol{X}_0 = \boldsymbol{x}] = \frac{\mathbb{E}\left[\prod_{i \in E} (\boldsymbol{X}_t)_i^{n_i} \mid \boldsymbol{X}_0 = \boldsymbol{x}\right]}{\mathbb{E}\left[\prod_{i \in E} (\boldsymbol{X}_\infty)_i^{n_i}\right]}. \tag{15}$$

If our random sample is interpreted as an independent and identically distributed (IID) set of n draws with replacement at time t from an infinite population evolving as a Wright-Fisher diffusion, then the quantity (15) is our claimed ratio of probabilities. Next, to interpret the right of (3), we must be able to assign a prior on $\boldsymbol{L}_0$. The appropriate choice is of course the sampling distribution of the coalescent, which can be shown under a PIM model

to be given by $m(\boldsymbol{n})$ in (12) (this is possible solely by coalescent arguments, without having to invoke the diffusion). Now, the right of (3) is

$$\mathbb{E}[F(\boldsymbol{x}, \boldsymbol{L}_t) \mid \boldsymbol{L}_0 = \boldsymbol{n}] = \mathbb{E}\left[\frac{\prod_{i\in E} x_i^{(\boldsymbol{L}_t)_i}}{m(\boldsymbol{L}_t)} \middle| \boldsymbol{L}_0 = \boldsymbol{n}\right].$$

The quantity inside the expectation is a ratio of: the probability of obtaining an ordered random sample with configuration $\boldsymbol{L}_t$ from a population in state $\boldsymbol{x}$ to the same probability under the coalescent prior. Two applications of Bayes' theorem then gives

$$\begin{aligned}\mathbb{E}\left[\frac{\prod_{i\in E} x_i^{(\boldsymbol{L}_t)_i}}{m(\boldsymbol{L}_t)} \middle| \boldsymbol{L}_0 = \boldsymbol{n}\right] &= \mathbb{E}\left[\frac{\mathbb{P}(\boldsymbol{X}_0 \in \mathrm{d}\boldsymbol{x} \mid \boldsymbol{L}_t)}{\mathbb{P}(\boldsymbol{X}_0 \in \mathrm{d}\boldsymbol{x})} \middle| \boldsymbol{L}_0 = \boldsymbol{n}\right] \\ &= \frac{\mathbb{P}(\boldsymbol{X}_0 \in \mathrm{d}\boldsymbol{x} \mid \boldsymbol{L}_0 = \boldsymbol{n})}{\mathbb{P}(\boldsymbol{X}_0 \in \mathrm{d}\boldsymbol{x})} = \frac{\mathbb{P}(\boldsymbol{L}_0 = \boldsymbol{n} \mid \boldsymbol{X}_0 = \boldsymbol{x})}{\mathbb{P}(\boldsymbol{L}_0 = \boldsymbol{n})}, \quad (16)\end{aligned}$$

which is again the claimed ratio (recalling that time 0 is different for $\boldsymbol{L}$ and $\boldsymbol{X}$). The right of (3) is therefore a ratio of *coalescent* sampling probabilities. The numerator is the probability for a random sample with configuration $\boldsymbol{n}$ given that the lineages ancestral to this sample a time t ago were typed by IID sampling from a population in state $\boldsymbol{x}$, while the denominator is the same probability without this additional information. Under this interpretation, the duality function (14) is also a ratio of sampling distributions, now without any offset of time:

$$F(\boldsymbol{x}, \boldsymbol{n}) = \frac{\mathbb{P}(\boldsymbol{L}_0 = \boldsymbol{n} \mid \boldsymbol{X}_t = \boldsymbol{x})}{\mathbb{P}(\boldsymbol{L}_0 = \boldsymbol{n})}.$$

3. An L-locus model

In this section we extend the above ideas to a multilocus model in which recombination can occur between each locus. We allow for more general mutation models than in Section 2, though for convenience we continue to assume that the type space at each locus is finite. We first introduce some notation. Suppose a haplotype is determined by the alleles at each of L loci. The set of possible alleles at locus l is denoted E_l, so that the set of all possible haplotypes is $E = \bigtimes_{l=1}^{L} E_l$. The frequency of haplotype $\boldsymbol{i} = (i_1, \ldots, i_L) \in E$ will be denoted by $x_{\boldsymbol{i}}$. The mutation parameter at locus l is θ_l and mutation

occurs at that locus according to a transition matrix $\boldsymbol{P}^{(l)} = (P^{(l)}_{ij})_{i,j\in E_l}$; in other words, when a mutation occurs to a haplotype with allele i at locus l, its offspring has allele j at that locus with probability $P^{(l)}_{ij}$. We will denote the resulting haplotype by $\boldsymbol{i}_{-l,j} := (i_1, \ldots, i_{l-1}, j, i_{l+1}, \ldots, i_L)$. Mutation occurs independently at each locus, so we may define mutation parameters across all loci as:

$$\theta = \sum_{l=1}^{L} \theta_l, \quad \boldsymbol{P} = \sum_{l=1}^{L} \frac{\theta_l}{\theta} \boldsymbol{I}_{|E_1|} \otimes \cdots \otimes \boldsymbol{I}_{|E_{l-1}|} \otimes \boldsymbol{P}^{(l)} \otimes \boldsymbol{I}_{|E_{l+1}|} \otimes \cdots \otimes \boldsymbol{I}_{|E_L|}, \quad (17)$$

where $\otimes$ denotes outer product, $\boldsymbol{I}_d$ is the $d \times d$ identity matrix, and $\boldsymbol{P}^{(l)}$ appears in the lth term in the product. Notice that if mutation is parent-independent at each locus (so $P^{(l)}_{ij} = P^{(l)}_j$ for each $l \in [L]$, $i, j \in E_l$), then the allele frequencies at each locus, $(X^{\{l\}}_{i_l})_{i_l \in E_l}$ with $X^{\{l\}}_{i_l} = \sum_{\boldsymbol{j}\in E:\, j_l = i_l} X_{\boldsymbol{j}}$, evolve marginally according to the one-locus model of Section 2. For each $l = 1, \ldots, L-1$, the rate of recombination between locus l and $l+1$ is parametrised by ρ_l, and we let $\rho = \sum_{l=1}^{L-1} \rho_l$.

For a nonempty subset $A \subseteq [L]$, denote the projection of E onto the co-ordinates in A by E_A, i.e. $E_A = \bigtimes_{l\in A} E_l$. Denote the marginal frequency of the alleles $\boldsymbol{i} \in E_A$ by

$$x^A_{\boldsymbol{i}} = \sum_{\boldsymbol{j}\in E:\, \boldsymbol{j}|_A = \boldsymbol{i}} x_{\boldsymbol{j}}.$$

Sometimes we will also write $x^A_{\boldsymbol{i}}$ for $\boldsymbol{i} \in E_B$ and $B \supset A$, by which it is implied that we mean

$$x^A_{\boldsymbol{i}|_A} = \sum_{\boldsymbol{j}\in E:\, \boldsymbol{j}|_A = \boldsymbol{i}|_A} x_{\boldsymbol{j}}. \quad (18)$$

Finally, for $A \subseteq [L]$ we also define the sets

$$A_{\leq l} = A \cap \{1, \ldots, l\}, \qquad A_{>l} = A \cap \{l+1, \ldots, L\}.$$

With this new definition for E, the multilocus Wright-Fisher diffusion process with recombination has state space Δ_E as in (5). Its generator is given by

$$\mathscr{L} = \frac{1}{2} \sum_{\boldsymbol{i}\in E} \left[\sum_{\boldsymbol{j}\in E} x_{\boldsymbol{i}}(\delta_{\boldsymbol{ij}} - x_{\boldsymbol{j}}) \frac{\partial}{\partial x_{\boldsymbol{j}}} \right.$$
$$\left. + \sum_{l=1}^{L} \theta_l \left[\sum_{j\in E_l} P^{(l)}_{ji_l} x_{\boldsymbol{i}_{-l,j}} - x_{\boldsymbol{i}} \right] + \sum_{l=1}^{L-1} \rho_l (x^{[L]_{\leq l}}_{\boldsymbol{i}} x^{[L]_{>l}}_{\boldsymbol{i}} - x_{\boldsymbol{i}}) \right] \frac{\partial}{\partial x_{\boldsymbol{i}}} \quad (19)$$

and $\mathscr{D}(\mathscr{L}) = C^2(\Delta_E)$.

3.1. An 'unreduced' dual

To obtain the dual process of (19), we follow the strategy outlined in Section 2. First consider the test function corresponding to the unordered sampling distribution of $\boldsymbol{n}$:

$$S(\boldsymbol{x}, \boldsymbol{n}) = \binom{n}{\boldsymbol{n}} \prod_{\boldsymbol{i} \in E} x_{\boldsymbol{i}}^{n_{\boldsymbol{i}}}, \tag{20}$$

where $\binom{n}{\boldsymbol{n}} = n! / \prod_{\boldsymbol{i} \in E} n_{\boldsymbol{i}}!$ is the multinomial coefficient. We know from Section 2 that, as a function of $\boldsymbol{x}$, our duality function will be proportional to $S(\boldsymbol{x}, \boldsymbol{n})$. In fact, rather than consider $S(\boldsymbol{x}, \boldsymbol{n})$ directly, we can work with the probability generating function (PGF)

$$G_n(\boldsymbol{s}; \boldsymbol{x}) = \sum_{\boldsymbol{n} \in \nabla_{E,n}} \left[\prod_{\boldsymbol{i} \in E} s_{\boldsymbol{i}}^{n_{\boldsymbol{i}}} \right] S(\boldsymbol{x}, \boldsymbol{n}) = \left[\sum_{\boldsymbol{i} \in E} s_{\boldsymbol{i}} x_{\boldsymbol{i}} \right]^n, \tag{21}$$

where $\boldsymbol{s} = (s_{\boldsymbol{i}})_{\boldsymbol{i} \in E}$ and

$$\nabla_{E,n} = \left\{ \boldsymbol{n} = (n_{\boldsymbol{i}})_{\boldsymbol{i} \in E} \in \mathbb{N}^{|E|} : \sum_{\boldsymbol{i} \in E} n_{\boldsymbol{i}} = n \right\},$$

and then recover $S(\boldsymbol{x}, \boldsymbol{n})$ from this later. (Here and throughout, define $S(\boldsymbol{x}, \boldsymbol{n}) = 0$ if $\boldsymbol{x} \notin \Delta_E$ or $\boldsymbol{n} \notin \nabla_{E,n}$ for any n.) For other examples of the use of generating functions in the context of population genetics models with recombination, see Griffiths (1981), Ethier and Griffiths (1990b), Griffiths (1991), and Lohse et al. (2011, 2016).

A simple calculation yields

$$\begin{aligned} \mathscr{L} G_n(\boldsymbol{s}; \boldsymbol{x}) = \sum_{\boldsymbol{i} \in E} \Bigg[\binom{n}{2} s_{\boldsymbol{i}}^2 x_{\boldsymbol{i}} G_{n-2}(\boldsymbol{s}; \boldsymbol{x}) + \sum_{l=1}^{L} \frac{\theta_l n}{2} \sum_{j \in E_l} s_{\boldsymbol{i}} P_{j i_l}^{(l)} x_{\boldsymbol{i}_{-l,j}} G_{n-1}(\boldsymbol{s}; \boldsymbol{x}) \\ + \sum_{l=1}^{L-1} \frac{\rho_l n}{2} s_{\boldsymbol{i}} x_{\boldsymbol{i}}^{[L] \leq l} x_{\boldsymbol{i}}^{[L] > l} G_{n-1}(\boldsymbol{s}; \boldsymbol{x}) \Bigg] - \frac{n(n-1+\theta+\rho)}{2} G_n(\boldsymbol{s}; \boldsymbol{x}). \end{aligned} \tag{22}$$

The remainder of the strategy would be (i) to extract the an equation for $\mathscr{L} S(\boldsymbol{x}, \boldsymbol{n})$ from (22), (ii) rearrange this equation into the required dual form, and (iii) read off a rate matrix for the dual process. However, we can see from

(20)–(22) that no distinction has been made between loci that are ancestral and those that are non-ancestral with respect to an 'initial' (present-day) sample. Consequently, the dual process would track both types of loci. This is the posterior analogue of the ARG of Griffiths and Marjoram (1997), in which the total number of lineages can grow unboundedly backwards in time. It would be preferable to construct an analogue of the 'reduced' version of the ARG in which only lineages ancestral to the initial sample are traced back in time (see, e.g. Hudson, 1983; Golding, 1984; Ethier and Griffiths, 1990b; Griffiths, 1991; Griffiths et al., 2008). We therefore move straight to the following subsection in which we construct a correspondingly reduced version of the dual process.

3.2. A 'reduced' dual

The state space for our reduced dual process will be

$$\Xi_{E,n} = \left\{ \boldsymbol{n} = (n_{\boldsymbol{i}}^A)_{\emptyset \neq A \subseteq [L], \boldsymbol{i} \in E_A} : n_{\boldsymbol{i}}^A \in \mathbb{N}, \sum_{\emptyset \neq A \subseteq [L]} \sum_{\boldsymbol{i} \in E_A} n_{\boldsymbol{i}}^A = n \right\}.$$

The set A records those loci at which the haplotype $\boldsymbol{i} \in E_A$ is ancestral to an initial (present-day) sample, and the alleles at only those loci are recorded. The notation $n_{\boldsymbol{i}}^A$ is then the number of times the haplotype $\boldsymbol{i}$ is observed, and we will also let $n^A = \sum_{\boldsymbol{i} \in E_A} n_{\boldsymbol{i}}^A$. By analogy with the previous subsection, we define the test function

$$\widetilde{S}(\boldsymbol{x}, \boldsymbol{n}) = \binom{n}{\boldsymbol{n}} \prod_{\emptyset \neq A \subseteq [L]} \prod_{\boldsymbol{i} \in E_A} (x_{\boldsymbol{i}}^A)^{n_{\boldsymbol{i}}^A}. \tag{23}$$

for $\boldsymbol{x} \in \Delta_E$, $\boldsymbol{n} \in \cup_{n=1}^{\infty} \Xi_{E,n}$ (and $\widetilde{S}(\boldsymbol{x}, \boldsymbol{n}) = 0$ otherwise); and the generating function

$$\begin{aligned} \widetilde{G}_n(\boldsymbol{t}; \boldsymbol{x}) &= \sum_{\boldsymbol{n} \in \Xi_{E,n}} \prod_{\emptyset \neq A \subseteq [L]} \prod_{\boldsymbol{i} \in E_A} (t_{\boldsymbol{i}}^A)^{n_{\boldsymbol{i}}^A} \widetilde{S}(\boldsymbol{x}, \boldsymbol{n}) \\ &= \left(\sum_{\emptyset \neq A \subseteq [L]} \sum_{\boldsymbol{i} \in E_A} t_{\boldsymbol{i}}^A x_{\boldsymbol{i}}^A \right)^n = \left[\sum_{\boldsymbol{j} \in E} \left(\sum_{\emptyset \neq A \subseteq [L]} t_{\boldsymbol{j}|_A}^A \right) x_{\boldsymbol{j}} \right]^n, \end{aligned} \tag{24}$$

with dummy variables $\boldsymbol{t} = (t_{\boldsymbol{i}}^A)_{\emptyset \neq A \subseteq [L], \boldsymbol{i} \in E_A}$, where the last equality follows from (18) and reordering the summations.

Now our use of generating functions pays off. Comparing the right-hand expression in (24) with (21) shows that to evaluate $\mathscr{L}\widetilde{G}_n(\boldsymbol{t};\boldsymbol{x})$ we simply need to apply the mapping

$$s_{\boldsymbol{i}} \mapsto \sum_{\emptyset\neq A\subseteq[L]} t^A_{\boldsymbol{i}|_A}$$

in (22). After some rearrangement we obtain

$$\begin{aligned}\mathscr{L}\widetilde{G}_n(\boldsymbol{t};\boldsymbol{x}) = \sum_{\emptyset\neq A\subseteq[L]} \Bigg[&\binom{n}{2} \sum_{\emptyset\neq B\subseteq[L]} \sum_{\boldsymbol{i}\in E_{A\cup B}} t^A_{\boldsymbol{i}} t^B_{\boldsymbol{i}} x^{A\cup B}_{\boldsymbol{i}} \widetilde{G}_{n-2}(\boldsymbol{t};\boldsymbol{x}) \\ &+ \sum_{\boldsymbol{i}\in E_A} \sum_{l\in A} \frac{\theta_l}{2} t^A_{\boldsymbol{i}} \sum_{j\in E_l} P^{(l)}_{ji_l} x^A_{\boldsymbol{i}_{-l,j}} \widetilde{G}_{n-1}(\boldsymbol{t};\boldsymbol{x}) \\ &+ \sum_{\boldsymbol{i}\in E_A} \sum_{l=\min A}^{\max A-1} \frac{\rho_l n}{2} t^A_{\boldsymbol{i}} x^{A_{\leq l}}_{\boldsymbol{i}} x^{A_{>l}}_{\boldsymbol{i}} \widetilde{G}_{n-1}(\boldsymbol{t};\boldsymbol{x}) \Bigg] \\ &- \left[\frac{n(n-1+\theta)}{2} + \sum_{l=1}^{L-1} \frac{\rho_l}{2} \sum_{\substack{A\subseteq[L]:\\ A_{\leq l}\neq\emptyset\neq A_{>l}}} n^A \right] \widetilde{G}_n(\boldsymbol{t};\boldsymbol{x}). \end{aligned} \tag{25}$$

Now we can continue the strategy outlined in the previous subsection. Noting that

$$\mathscr{L}\widetilde{G}_n(\boldsymbol{t};\boldsymbol{x}) = \sum_{\boldsymbol{n}\in\Xi_{E,n}} \prod_{\emptyset\neq A\subseteq[L]} \prod_{\boldsymbol{i}\in E_A} (t^A_{\boldsymbol{i}})^{n^A_{\boldsymbol{i}}} \mathscr{L}\widetilde{S}(\boldsymbol{x},\boldsymbol{n}),$$

we can compare coefficients of $\prod_{\emptyset\neq A\subseteq[L]} \prod_{\boldsymbol{i}\in E_A} (t^A_{\boldsymbol{i}})^{n^A_{\boldsymbol{i}}}$ in (25) to obtain

$$\begin{aligned}\mathscr{L}\widetilde{S}(\boldsymbol{x},\boldsymbol{n}) = \sum_{\emptyset\neq A\subseteq[L]} \Bigg[&\binom{n}{2} \sum_{\emptyset\neq B\subseteq[L]} \sum_{\boldsymbol{i}\in E_{A\cup B}} x^{A\cup B}_{\boldsymbol{i}} \widetilde{S}(\boldsymbol{x},\boldsymbol{n}-\boldsymbol{e}^A_{\boldsymbol{i}}-\boldsymbol{e}^B_{\boldsymbol{i}}) \\ &+ \sum_{\boldsymbol{i}\in E_A} \sum_{l\in A} \frac{\theta_l}{2} \sum_{j\in E_l} P^{(l)}_{ji_l} x^A_{\boldsymbol{i}_{-l,j}} \widetilde{S}(\boldsymbol{x},\boldsymbol{n}-\boldsymbol{e}^A_{\boldsymbol{i}}) \\ &+ \sum_{\boldsymbol{i}\in E_A} \sum_{l=\min A}^{\max A-1} \frac{\rho_l n}{2} x^{A_{\leq l}}_{\boldsymbol{i}} x^{A_{>l}}_{\boldsymbol{i}} \widetilde{S}(\boldsymbol{x},\boldsymbol{n}-\boldsymbol{e}^A_{\boldsymbol{i}}) \Bigg] \end{aligned}$$

$$- \left[\frac{n(n-1+\theta)}{2} + \sum_{l=1}^{L-1} \frac{\rho_l}{2} \sum_{\substack{A\subseteq[L]:\\ A_{\le l}\neq\emptyset\neq A_{>l}}} n^A \right] \widetilde{S}(\boldsymbol{x},\boldsymbol{n}). \qquad (26)$$

To manipulate this into dual form, we further rearrange the right-hand side in order to remove the explicit instances of $\boldsymbol{x}$ outside of $S(\boldsymbol{x},\cdot)$. Using (A.1)–(A.3) of Appendix A together with (26), we obtain

$$\begin{aligned}
\mathscr{L}\widetilde{S}(\boldsymbol{x},\boldsymbol{n}) = \\
&\frac{1}{2}\sum_{\emptyset\neq A\subseteq[L]} \Bigg[\sum_{\emptyset\neq B\subseteq[L]} \sum_{\boldsymbol{i}\in E_{A\cup B}} n(n_{\boldsymbol{i}}^{A\cup B} + 1 - \delta_{A,A\cup B} - \delta_{B,A\cup B}) \\
&\qquad\qquad \times \widetilde{S}(\boldsymbol{x}, \boldsymbol{n} - \boldsymbol{e}_{\boldsymbol{i}}^A - \boldsymbol{e}_{\boldsymbol{i}}^B + \boldsymbol{e}_{\boldsymbol{i}}^{A\cup B}) \\
&+ \sum_{\boldsymbol{i}\in E_A} \sum_{l\in A} \theta_l \sum_{j\in E_l} P_{ji_l}^{(l)} (n_{\boldsymbol{i}_{-l,j}}^A + 1 - \delta_{i_l j}) \widetilde{S}(\boldsymbol{x}, \boldsymbol{n} - \boldsymbol{e}_{\boldsymbol{i}}^A + \boldsymbol{e}_{\boldsymbol{i}_{-l,j}}^A) \\
&+ \sum_{\boldsymbol{i}\in E_A} \sum_{l=\min A}^{\max A-1} \rho_l \frac{(n_{\boldsymbol{i}}^{A_{\le l}}+1)(n_{\boldsymbol{i}}^{A_{>l}}+1)}{n+1} \widetilde{S}(\boldsymbol{x}, \boldsymbol{n} - \boldsymbol{e}_{\boldsymbol{i}}^A + \boldsymbol{e}_{\boldsymbol{i}}^{A_{\le l}} + \boldsymbol{e}_{\boldsymbol{i}}^{A_{>l}}) \Bigg] \\
&- \left[\frac{n(n+\theta-1)}{2} + \sum_{l=1}^{L-1} \frac{\rho_l}{2} \sum_{\substack{A\subseteq[L]:\\ A_{\le l}\neq\emptyset\neq A_{>l}}} n^A \right] \widetilde{S}(\boldsymbol{x},\boldsymbol{n}). \qquad (27)
\end{aligned}$$

If we divide (27) by $\mathbb{E}[\widetilde{S}(\boldsymbol{X}_\infty,\boldsymbol{n})]$ then, after a little rearrangement, we have succeeded in writing $\mathscr{L}F(\boldsymbol{x},\boldsymbol{n})$ in the form of (9) for the duality function

$$\widetilde{F}(\boldsymbol{x},\boldsymbol{n}) = \frac{\widetilde{S}(\boldsymbol{x},\boldsymbol{n})}{\mathbb{E}[\widetilde{S}(\boldsymbol{X}_\infty,\boldsymbol{n})]}, \qquad (28)$$

from which we can read off the rate matrix for the dual process on $\cup_{n=1}^\infty \Xi_{E,n}$. We have therefore shown the following.

Theorem 1. *Let*

$$\widetilde{m}(\boldsymbol{n}) = \mathbb{E}\left[\prod_{\emptyset\neq A\subseteq[L]} \prod_{\boldsymbol{i}\in E_A} (X_{\boldsymbol{i}}^A)^{n_{\boldsymbol{i}}^A} \right] \qquad (29)$$

(for $\boldsymbol{n} \in \cup_{n=1}^{\infty} \Xi_{E,n}$ and 0 otherwise), where expectation is taken with respect to the stationary distribution of $\boldsymbol{X}$. The Wright-Fisher diffusion $\boldsymbol{X} = (\boldsymbol{X}_t)_{t\geq 0}$ on Δ_E with generator (19) is dual to a pure jump process $\widetilde{\boldsymbol{L}} = (\widetilde{\boldsymbol{L}}_t)_{t\geq 0}$ on $\cup_{n=1}^{\infty} \Xi_{E,n}$ with transitions given by the following description.

Coalescence. *For each nonempty $A, B \subseteq [L]$ and each $\boldsymbol{i} \in E_{A\cup B}$, the process jumps to $\boldsymbol{n} - \boldsymbol{e}_{\boldsymbol{i}}^{A} - \boldsymbol{e}_{\boldsymbol{i}}^{B} + \boldsymbol{e}_{\boldsymbol{i}}^{A\cup B}$ at rate*

$$\frac{1}{2}\frac{\widetilde{m}(\boldsymbol{n} - \boldsymbol{e}_{\boldsymbol{i}}^{A} - \boldsymbol{e}_{\boldsymbol{i}}^{B} + \boldsymbol{e}_{\boldsymbol{i}}^{A\cup B})}{\widetilde{m}(\boldsymbol{n})} n_{\boldsymbol{i}}^{A}(n_{\boldsymbol{i}}^{B} - \delta_{AB}).$$

Mutation. *For each nonempty $A \subseteq [L]$, $l \in A$, $\boldsymbol{i} \in E_A$, and $j \in E_l$, the process jumps to $\boldsymbol{n} - \boldsymbol{e}_{\boldsymbol{i}}^{A} + \boldsymbol{e}_{\boldsymbol{i}_{-l,j}}^{A}$ at rate*

$$\frac{1}{2}\frac{\widetilde{m}(\boldsymbol{n} - \boldsymbol{e}_{\boldsymbol{i}}^{A} + \boldsymbol{e}_{\boldsymbol{i}_{-l,j}}^{A})}{\widetilde{m}(\boldsymbol{n})} n_{\boldsymbol{i}}^{A}\theta_l P_{ji_l}^{(l)}.$$

Recombination. *For each nonempty $A \subseteq [L]$, $\boldsymbol{i} \in E_A$, and $l = \min A, \ldots, \max A - 1$, the process jumps to $\boldsymbol{n} - \boldsymbol{e}_{\boldsymbol{i}}^{A} + \boldsymbol{e}_{\boldsymbol{i}}^{A_{\leq l}} + \boldsymbol{e}_{\boldsymbol{i}}^{A_{>l}}$ at rate*

$$\frac{1}{2}\frac{\widetilde{m}(\boldsymbol{n} - \boldsymbol{e}_{\boldsymbol{i}}^{A} + \boldsymbol{e}_{\boldsymbol{i}}^{A_{\leq l}} + \boldsymbol{e}_{\boldsymbol{i}}^{A_{>l}})}{\widetilde{m}(\boldsymbol{n})} n_{\boldsymbol{i}}^{A}\rho_l.$$

The duality function relating the two processes is $\widetilde{F}(\boldsymbol{x}, \boldsymbol{n})$, given by (28) and (23).

Remark 1. *It is straightforward, though notationally cumbersome, to construct $\boldsymbol{Q}$ from the description given in Theorem 1. A given transition rate $q(\boldsymbol{n}, \widehat{\boldsymbol{n}})$ is obtained by summing over the rates in Theorem 1 that correspond to a particular destination state $\widehat{\boldsymbol{n}}$.*

Corollary 1. *The transient sampling distributions of $\boldsymbol{X}$ and $\widetilde{\boldsymbol{L}}$ are related by*

$$\mathbb{E}[\widetilde{S}(\boldsymbol{X}_t, \boldsymbol{n}) \mid \boldsymbol{X}_0 = \boldsymbol{x}] = \mathbb{E}\left[\frac{\mathbb{E}[\widetilde{S}(\boldsymbol{X}_\infty, \boldsymbol{n})]}{\mathbb{E}[\widetilde{S}(\boldsymbol{X}_\infty, \widetilde{\boldsymbol{L}}_t) \mid \widetilde{\boldsymbol{L}}_t]}\widetilde{S}(\boldsymbol{x}, \widetilde{\boldsymbol{L}}_t) \mid \widetilde{\boldsymbol{L}}_0 = \boldsymbol{n}\right].$$

Proof. This follows immediately from the duality equation

$$\mathbb{E}\left[\widetilde{F}(\boldsymbol{X}_t, \boldsymbol{n}) \mid \boldsymbol{X}_0 = \boldsymbol{x}\right] = \mathbb{E}\left[\widetilde{F}(\boldsymbol{x}, \widetilde{\boldsymbol{L}}_t) \mid \widetilde{\boldsymbol{L}}_0 = \boldsymbol{n}\right]$$

and (28). □

It is possible to provide a genealogical interpretation of Theorem 1 in a spirit similar to that given in Section 2, the main differences being that here we account for recombination between multiple loci and construct the dual process so that it tracks only lineages ancestral to the initial sample. In summary, the duality function (28) is proportional to the ordered sampling distribution $\prod_{\emptyset\neq A\subseteq[L]}\prod_{\boldsymbol{i}\in E_A}(X_{\boldsymbol{i}}^A)^{n_{\boldsymbol{i}}^A}$ of a haplotype configuration $\boldsymbol{n}$, when sampling is performed IID from a population with haplotype frequencies $\boldsymbol{x}$. In this interpretation, the set of loci A at which a sampled haplotype is actually observed is nonrandom. The normalisation constant of (28) is then $\widetilde{m}(\boldsymbol{n})$, the sampling distribution for $\boldsymbol{n}$ when the population haplotype frequencies are at stationarity; this ensures both that equation (27) can easily be identified as the generator of a process acting on $\boldsymbol{n}$, and that the duality equation (3) has a straightforward interpretation as two ways of looking at (a ratio of) sampling probabilities. In this duality equation it is necessary to consider the genealogy of the present-day configuration $\boldsymbol{n}$ conditioned on the past state of the population, which gives rise to the posterior coalescent dynamics captured by the process $\widetilde{\boldsymbol{L}}$ described in Theorem 1. The ratios of terms in $\widetilde{m}(\cdot)$ in the transition rates of $\widetilde{\boldsymbol{L}}$ appear naturally as a time-reversal of the coalescent process.

Of course, a major complication of the dual process here compared to that of Section 2 is that there is no closed-form expression for the stationary moments $\widetilde{m}(\boldsymbol{n})$ of $\boldsymbol{X}$ [eq. (29)]. However, we can show that they satisfy a simple linear system.

Proposition 1. *For $\boldsymbol{n}\in\Xi_{E,n}$, the stationary moments $\widetilde{m}(\boldsymbol{n})$ of* (29) *satisfy*

$$\begin{aligned}
&\left[n(n-1)+\sum_{l=1}^{L}\theta_l\sum_{\substack{A\subseteq[L]:\\ l\in A}}n^A+\sum_{l=1}^{L-1}\rho_l\sum_{\substack{A\subseteq[L]:\\ A_{\leq l}\neq\emptyset\neq A_{>l}}}n^A\right]\widetilde{m}(\boldsymbol{n})=\\
&\qquad\sum_{\emptyset\neq A\subseteq[L]}\left[\sum_{\emptyset\neq B\subseteq[L]}\sum_{\boldsymbol{i}\in E_{A\cup B}}n_{\boldsymbol{i}}^A(n_{\boldsymbol{i}}^B-\delta_{AB})\widetilde{m}(\boldsymbol{n}-\boldsymbol{e}_{\boldsymbol{i}}^A-\boldsymbol{e}_{\boldsymbol{i}}^B+\boldsymbol{e}_{\boldsymbol{i}}^{A\cup B})\right.\\
&\qquad+\sum_{\boldsymbol{i}\in E_A}\sum_{l\in A}\theta_l\sum_{j\in E_A}P^{(l)}_{ji_l}n_{\boldsymbol{i}}^A\widetilde{m}(\boldsymbol{n}-\boldsymbol{e}_{\boldsymbol{i}}^A+\boldsymbol{e}_{\boldsymbol{i}_{-l,j}}^A)\\
&\qquad\left.+\sum_{\boldsymbol{i}\in E_A}\sum_{l=\min A}^{\max A-1}\rho_l n_{\boldsymbol{i}}^A\widetilde{m}(\boldsymbol{n}-\boldsymbol{e}_{\boldsymbol{i}}^A+\boldsymbol{e}_{\boldsymbol{i}}^{A_{\leq l}}+\boldsymbol{e}_{\boldsymbol{i}}^{A_{>l}})\right].
\end{aligned}\tag{30}$$

A boundary condition is

$$\widetilde{m}(\boldsymbol{e}_{i_1}^{[1]} + \boldsymbol{e}_{i_2}^{[2]} + \cdots + \boldsymbol{e}_{i_L}^{[L]}) = \prod_{l=1}^{L} \mu_{i_l}^{(l)}, \qquad i_l \in E_l,$$

where $\boldsymbol{\mu}^{(l)} = (\mu_1^{(l)}, \mu_2^{(l)}, \ldots, \mu_{|E_l|}^{(l)})$ *is the stationary distribution of* $\boldsymbol{P}^{(l)}$.

Proof. Take expectation with respect to the stationary distribution of $\boldsymbol{X}$ in (27) and apply the identity $\mathbb{E}[\mathscr{L}\widetilde{S}(\boldsymbol{X}_\infty, \boldsymbol{n})] = 0$ to get (30). The boundary condition follows by the argument of Fearnhead (2003, Theorem 1). □

The advantage of a *reduced* dual is now apparent. If we define the *degree* of $\boldsymbol{n}$ by

$$\text{degree}(\boldsymbol{n}) = \sum_{\emptyset \neq A \subseteq [L]} |A| n^A,$$

the total length of all ancestral material in the sample, then the system (30) is *closed* in the sense that terms on the right of (30) have degree less than or equal to that of $\boldsymbol{n}$, and so it can in principle be solved (e.g. by matrix inversion). The process $\widetilde{\boldsymbol{L}}$ evolves on a *finite* state space. This is not true of the unreduced dual.

Recursive systems similar to (30) have been studied by Griffiths (1981), Golding (1984), Ethier and Griffiths (1990b), Larribe et al. (2002), Fearnhead (2003), Griffiths et al. (2008), Jenkins and Song (2009), Larribe and Lessard (2008), and Jenkins and Griffiths (2011), among others. With the exception of Larribe and Lessard (2008), whose eq. (1) is equal to (30) up to a combinatorial factor, typically these studies focus on special cases such as two loci or parameters uniform across loci. It is common in studying systems of this form to derive them by a probabilistic argument; in particular, by describing the associated coalescent process and partitioning on each of the most recent possible events going back in time. This approach can be combinatorially involving, and we emphasise the cleanliness of the method taken in this paper: once we have the generator (19), the rest follows mechanistically.

3.3. A closed-form solution

One special case of the above model permits a closed-form solution: mutation within each locus is parent-independent ($P_{ij}^{(l)} = P_j^{(l)}$ for each $l \in [L]$,

$i, j \in E_l$), and $\rho_l = \infty$ for each $l \in [L]$. Then each locus evolves independently, and the dual process is projected onto the subspace $\Xi^{\infty}_{E,n} \subseteq \Xi_{E,n}$ given by

$$\Xi^{\infty}_{E,n} = \left\{ \boldsymbol{n} \in \Xi_{E,n} : n^A = 0 \; \forall A \notin \{\{1\}, \{2\}, \ldots, \{L\}\} \right\};$$

that is, one in which any haplotype is ancestral at precisely one locus. The projection is achieved by mapping a haplotype $\boldsymbol{i} \in E_A$ with $A = \{a_1, \ldots, a_{|A|}\}$ to $|A|$ different haplotypes of type $i_1 \in E_{a_1}$, $i_2 \in E_{a_2}, \ldots, i_{|A|} \in E_{a_{|A|}}$; recombination instantaneously breaks apart each locus. The generator for this model is a sum of L generators acting independently on each locus (see Ethier and Griffiths, 1990a, for further details), from which we can write down the transition rates of the dual process on $\cup_{m=1}^{Ln} \Xi^{\infty}_{E,n}$:

$$\widetilde{q}(\boldsymbol{n}, \widehat{\boldsymbol{n}}) = \frac{1}{2} \times \begin{cases} n_i^{\{l\}}(n^{\{l\}} + \theta_l - 1) & \text{if } \widehat{\boldsymbol{n}} = \boldsymbol{n} - \boldsymbol{e}_i^{\{l\}} \\ & \text{where } l \in [L], i \in E_l, \\ -\sum_{l=1}^{L} n^{\{l\}}(n^{\{l\}} + \theta_l - 1) & \text{if } \widehat{\boldsymbol{n}} = \boldsymbol{n}. \end{cases}$$

The duality function in this case is

$$\widetilde{F}(\boldsymbol{x}, \boldsymbol{n}) = \prod_{l=1}^{L} \left[\frac{(\theta_l)_{n^{\{l\}}}}{\prod_{i \in E_l} (\theta_l P_i^{(l)})_{n_i^{\{l\}}}} \prod_{i \in E_l} (x_i^{\{l\}})^{n_i^{\{l\}}} \right],$$

which is simply the product of L copies of the one-locus duality function encountered in Section 2, as it must be under free recombination.

4. A transition function expansion

Duality can be used to obtain an expression for the transition function of the Wright-Fisher diffusion. Here we tackle the diffusion with generator (19), whose transition density with respect to Lebesgue measure, after evolving from $\boldsymbol{x}$ for a time t, we denote by $f(\boldsymbol{x}, \cdot; t)$, and whose stationary distribution we denote by $\pi(\cdot)$. To our knowledge this is the first time an expression for the transition function of a Wright-Fisher diffusion has incorporated recombination.

For simplicity we restrict our attention to 'completely specified' samples: those for which $n_{\boldsymbol{i}}^A = 0$ if $A \neq [L]$, and we write $n_{\boldsymbol{i}}$ for $n_{\boldsymbol{i}}^{[L]}$, and so on. Then the sampling distribution of $\boldsymbol{n}$ can be written in the simpler form of (20).

Our result will be expressed in terms of the transitions of the dual process, which we denote $p_{\boldsymbol{nl}}(t) := \mathbb{P}(\widetilde{\boldsymbol{L}}_t = \boldsymbol{l} \mid \widetilde{\boldsymbol{L}}_0 = \boldsymbol{n})$. In particular, we let $n \to \infty$ in such a way that $\boldsymbol{n}/n \to \boldsymbol{y} \in \Delta_E$ (this idea is formalised by Barbour et al., 2000, p125) and write

$$p_{\boldsymbol{yl}}(t) := \lim_{n\to\infty:\frac{\boldsymbol{n}}{n}\to\boldsymbol{y}} p_{\boldsymbol{nl}}(t). \tag{31}$$

The existence of this limit ensures that our typed, reduced, coalescent process $\widetilde{\boldsymbol{L}}$ can be initiated from infinitely many lineages.

Theorem 2. *Suppose that* (31) *defines a probability distribution on* $\bigcup_{n=1}^{\infty} \Xi_{E,n}$ *for each* $t > 0$, $\boldsymbol{y} \in \Delta_E$. *Then the transition density function of the Wright-Fisher diffusion with generator* (19) *is given by*

$$f(\boldsymbol{x}, \boldsymbol{y}; t) = \pi(\boldsymbol{y}) \sum_{\boldsymbol{l}\in\bigcup_{l\in\mathbb{N}}\Xi_{E,l}} \frac{p_{\boldsymbol{yl}}(t)}{\widetilde{m}(\boldsymbol{l})} \prod_{\emptyset\neq A\subseteq[L]} \prod_{\boldsymbol{i}\in E_A} (x_{\boldsymbol{i}}^A)^{l_{\boldsymbol{i}}^A}, \tag{32}$$

with $\widetilde{m}(\cdot)$ *as in* (29).

Proof. The proof is similar to the rigorous treatment given in Barbour et al. (2000) and so we give only a summary. Corollary 1 easily leads to

$$\begin{aligned}\mathbb{E}\left[S(\boldsymbol{X}_t, \boldsymbol{n}) \middle| \boldsymbol{X}_0 = \boldsymbol{x}\right] &= \binom{n}{\boldsymbol{n}} m(\boldsymbol{n}) \mathbb{E}\left[\frac{\prod_{\emptyset\neq A\subseteq[L]}\prod_{\boldsymbol{i}\in E_A}(x_{\boldsymbol{i}}^A)^{\widetilde{L}_{\boldsymbol{i}}^A(t)}}{\widetilde{m}(\widetilde{\boldsymbol{L}}_t)} \middle| \widetilde{\boldsymbol{L}}_0 = \boldsymbol{n}\right] \\ &= \binom{n}{\boldsymbol{n}} m(\boldsymbol{n}) \sum_{\boldsymbol{l}\in\bigcup_{l\leq Ln}\Xi_{E,l}} \frac{p_{\boldsymbol{nl}}(t)}{\widetilde{m}(\boldsymbol{l})} \prod_{\emptyset\neq A\subseteq[L]} \prod_{\boldsymbol{i}\in E_A} (x_{\boldsymbol{i}}^A)^{l_{\boldsymbol{i}}^A}. \end{aligned} \tag{33}$$

Our aim is to let $n \to \infty$ and $\boldsymbol{n}/n \to \boldsymbol{y}$ in (33). Letting $n \to \infty$ on the left-hand side is tantamount to identifying a distribution from its moments. Etheridge and Griffiths (2009) note that this is an application of 'sample inversion': for a continuous function $u : \Delta_E \to \mathbb{R}$ and a random sample $\boldsymbol{N} \sim \text{Multinomial}(n, \boldsymbol{z})$,

$$\mathbb{E}\left[u\left(\frac{\boldsymbol{N}}{n}\right)\right] = \sum_{\boldsymbol{k}\in\nabla_{E,n}} u\left(\frac{\boldsymbol{k}}{n}\right) S(\boldsymbol{z}, \boldsymbol{k}) \to u(\boldsymbol{z}), \qquad n \to \infty,$$

uniformly in $\boldsymbol{z} \in \Delta_E$. To use this result we multiply both sides of (33) by $u(\boldsymbol{n}/n)$. If u is a function such that $u(\boldsymbol{k}/n) = 0$ if $\boldsymbol{k} \neq \boldsymbol{n}$ then the left-hand side of the resulting equation is

$$\begin{aligned}\mathbb{E}\left[u(\boldsymbol{n}/n)S(\boldsymbol{X}_t, \boldsymbol{n})|\, \boldsymbol{X}_0 = \boldsymbol{x}\right] &= \mathbb{E}\left[\mathbb{E}[u(\boldsymbol{N}/n)|\, \boldsymbol{X}_t]|\, \boldsymbol{X}_0 = \boldsymbol{x}\right] \\ &\to \mathbb{E}[u(\boldsymbol{X}_t)|\boldsymbol{X}_0 = \boldsymbol{x}]\end{aligned}$$

as $n \to \infty$, where $\boldsymbol{N} \sim \text{Multinomial}(n, \boldsymbol{X}_t)$ and the interchange of limit and integral is justified by Barbour et al. (2000). Similarly,

$$\sum_{\boldsymbol{k} \in \nabla_{E,n}} \binom{n}{\boldsymbol{n}} m(\boldsymbol{n}) u\left(\frac{\boldsymbol{k}}{n}\right) \to \mathbb{E}[u(\boldsymbol{X}_\infty)], \qquad n \to \infty.$$

These arguments can be shown still to hold if u is replaced by a delta function at $\boldsymbol{y}$ (Barbour et al., 2000), and then $\mathbb{E}[u(\boldsymbol{X}_t)|\, \boldsymbol{X}_0 = \boldsymbol{x}] = f(\boldsymbol{x}, \boldsymbol{y}; t)$ and $\mathbb{E}[u(\boldsymbol{X}_\infty)] = \pi(\boldsymbol{y})$. Put all this together and let $n \to \infty$ to yield (32). □

Equation (32) has an intuitive interpretation via Bayes' theorem (Figure 1), similar to the one given in Section 2. The conditional density of $\boldsymbol{y}|\boldsymbol{x}$ is proportional to its prior density $\pi(\boldsymbol{y})$ times the conditional density of $\boldsymbol{x}|\boldsymbol{y}$. The information that $\boldsymbol{y}$ transfers to the conditional density of $\boldsymbol{x}$ flows through the dual process $\widetilde{\boldsymbol{L}}$, which evolves back from an initial state $\boldsymbol{y}$ to a state $\boldsymbol{l}$ a time t ago (with probability $p_{\boldsymbol{yl}}(t)$). Given $\widetilde{\boldsymbol{L}}_t = \boldsymbol{l}$, the density of $\boldsymbol{x}$ is proportional to the likelihood of the type configuration associated with $\boldsymbol{l}$ given $\boldsymbol{x}$ (contributing the multinomial term). The normalisation of this conditional likelihood is the marginal likelihood $\widetilde{m}(\boldsymbol{l})$ of $\boldsymbol{l}$ when $\boldsymbol{x}$ is integrated over its (prior) stationary distribution.

Remark 2. *For reversible diffusions one can obtain a version of the transition density more flexible than* (32)*, expressed in terms of $p_{\boldsymbol{xl}}(t)$ rather than $p_{\boldsymbol{yl}}(t)$. Despite the interchange of $\boldsymbol{x}$ and $\boldsymbol{y}$, it is still possible to interpret the alternative form for the transition density in terms of a dual process running backwards in time (Donnelly and Tavaré, 1987; Etheridge and Griffiths, 2009). However, the Wright-Fisher diffusion with recombination is not reversible (Handa, 2002).*

Remark 3. *The existence of $p_{\boldsymbol{yl}}(t)$ in a model incorporating selection rather than recombination is proven rigorously by Barbour et al. (2000). It may be possible to adapt their approach here; we leave this for future work.*

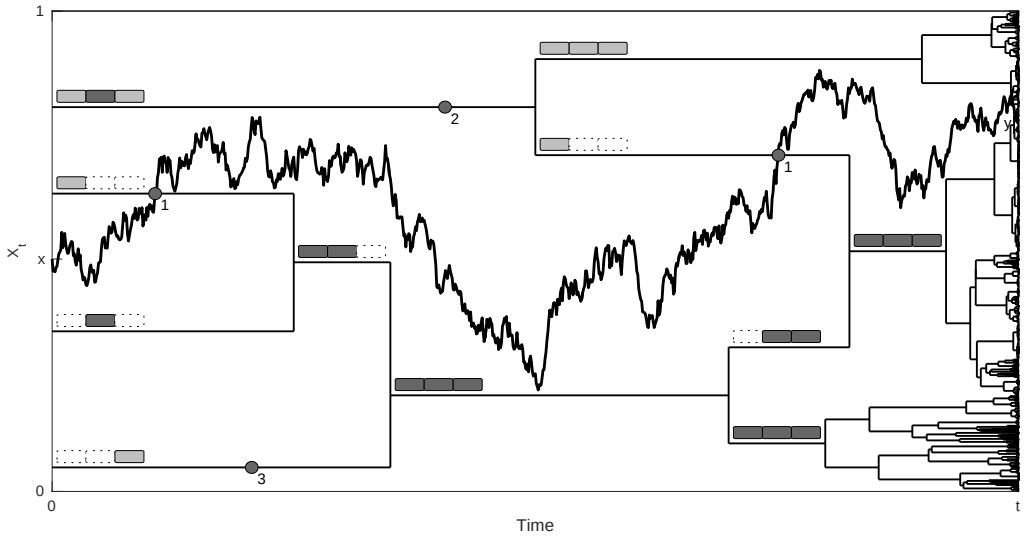


Figure 1: Illustration of the transition density in an $L = 3$ locus model, with two alleles at each locus (shown in dark and light grey). The diffusion $\boldsymbol{X}$ evolves from $\boldsymbol{X}_0 = \boldsymbol{x}$ to $\boldsymbol{X}_t = \boldsymbol{y}$ (only the first co-ordinate is plotted). The dual jump process, $\widetilde{\boldsymbol{L}}$, shown here as a typed ARG, evolves back in time from infinitely many lineages at time t with configuration $\widetilde{\boldsymbol{L}}_0 = \boldsymbol{y}$ to a configuration $\widetilde{\boldsymbol{L}}_t$ of size 4 at time 0 (note the time index now runs backwards). The haplotype associated with each lineage is shown as three shaded segments, and non-ancestral loci are shown with a dotted outline. Mutations in the graph are shown as circles and are labelled by the locus they affect. Denoting the dark and light alleles by 0 and 1 respectively, the four types $(\boldsymbol{i}, A)$ of $\widetilde{\boldsymbol{L}}_t$ are, from top to bottom, $((1,0,1),\{1,2,3\}), ((1),\{1\}), ((0),\{2\}), ((1),\{3\})$.

5. A continuous model

Before studying the L-locus model further, we illustrate how the above strategy can also be applied to a continuous model of recombination. For this to make sense the mutation model should also be continuous, and an appropriate choice is the infinitely-many-sites model. One way to achieve the appropriate duality result is first to write down the relevant diffusion model and then to pursue the strategy above, for example by recasting it as a Fleming-Viot measure-valued diffusion along the lines of Ethier and Griffiths (1987). Here we take a more direct approach by taking the formal limit in the L-locus model as $L \to \infty$. To take this limit painlessly we will reformulate our L-locus model somewhat.

First consider a representation for the continuous limit. Here a chro-

mosome is idealised as the interval $[0,1]$, and the model is specified by two probability measures on $[0,1]$, which we assume to admit densities η and ν with respect to Lebesgue measure, respectively modelling the distribution of mutation and recombination events along a chromosome. (The usual infinitely-many-sites model of mutation is recovered by letting $\eta(x) \equiv 1$. This is also a typical choice for ν.) A haplotype in this model can be specified by a set $\boldsymbol{\xi} \subseteq [0,1]$ of positions at which it differs from some reference haplotype. If the reference haplotype is chosen to be that of the grand most recent common ancestor of a sample of n haplotypes, then $|\boldsymbol{\xi}|$ is finite (Griffiths and Marjoram, 1997). The state space for this model is

$$\Xi_{[0,1],n} := \left\{ \boldsymbol{n} = (n_{\boldsymbol{i}}^A)_{\emptyset \neq A \subseteq [0,1], \boldsymbol{\xi} \subseteq A} : |\boldsymbol{\xi}| < \infty, n_{\boldsymbol{\xi}}^A \in \mathbb{N}, \sum_{\emptyset \neq A \subseteq [0,1]} \sum_{\boldsymbol{\xi} \subseteq A} n_{\boldsymbol{\xi}}^A = n \right\},$$

with each A Borel measurable.

We embed the L-locus model in this continuous description by the mapping $[L] \mapsto \left\{\frac{1}{L}, \frac{2}{L}, \ldots, 1\right\}$. Then a mutation at locus l, or a recombination between locus l and $l+1$, occurs at position l/L, and we choose

$$E_l = \{1,2\}, \quad \boldsymbol{P}^{(l)} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \theta_l = \theta \int_{\frac{l-1}{L}}^{\frac{l}{L}} \nu(x)\mathrm{d}x, \quad \rho_l = \rho \int_{\frac{l-1}{L}}^{\frac{l}{L}} \eta(x)\mathrm{d}x,$$

for each $l \in [L]$. In Appendix B we show that if we let $L \to \infty$ then this embedding recovers a well-defined limiting process for the dual, with state space $\Xi_{[0,1],n}$, and with a mixture of diffuse and atomic jump kernels. It can be described as follows. Given that the process is currently in state $\boldsymbol{n} \in \Xi_{[0,1],n}$:

Coalescence. For each $A, B \subseteq [0,1]$ and $\boldsymbol{\xi} \subseteq A \cup B$, the process jumps to $\boldsymbol{n} - \boldsymbol{e}_{\boldsymbol{\xi}}^A - \boldsymbol{e}_{\boldsymbol{\xi}}^B + \boldsymbol{e}_{\boldsymbol{\xi}}^{A \cup B}$ at rate

$$\frac{1}{2} n_{\boldsymbol{\xi}}^A (n_{\boldsymbol{\xi}}^B - \delta_{AB}) \frac{\widetilde{m}(\boldsymbol{n} - \boldsymbol{e}_{\boldsymbol{\xi}}^A - \boldsymbol{e}_{\boldsymbol{\xi}}^B + \boldsymbol{e}_{\boldsymbol{\xi}}^{A \cup B})}{\widetilde{m}(\boldsymbol{n})}.$$

Mutation. For each $A \subseteq [0,1]$ and $\boldsymbol{\xi} \subseteq A$, the process jumps at rate

$$\frac{\theta}{2} n_{\boldsymbol{\xi}}^A \int_A \frac{\widetilde{m}(\boldsymbol{n} - \boldsymbol{e}_{\boldsymbol{\xi}}^A + \boldsymbol{e}_{\boldsymbol{\xi}(x)}^A)}{\widetilde{m}(\boldsymbol{n})} \eta(x)\mathrm{d}x, \tag{34}$$

where

$$\bar{\boldsymbol{\xi}}(x) = \begin{cases} \boldsymbol{\xi} \setminus \{x\} & \text{if } x \in \boldsymbol{\xi}, \\ \boldsymbol{\xi} \cup \{x\} & \text{if } x \notin \boldsymbol{\xi}. \end{cases} \tag{35}$$

The resulting state is $\boldsymbol{n} - \boldsymbol{e}^A_{\boldsymbol{\xi}} + \boldsymbol{e}_{\boldsymbol{\xi}\setminus\{x\}}$, where the position $x \in \boldsymbol{\xi}$ is chosen by the probability distribution proportional to $\frac{\widetilde{m}(\boldsymbol{n} - \boldsymbol{e}^A_{\boldsymbol{\xi}} + \boldsymbol{e}^A_{\boldsymbol{\xi}\setminus\{x\}})}{\widetilde{m}(\boldsymbol{n})}\eta(x)\mathrm{d}x$.

Recombination. For each $A \subseteq [0,1]$ and $\boldsymbol{\xi} \subseteq A$, the process jumps at rate

$$\frac{\rho}{2} n^A_{\boldsymbol{\xi}} \int_{\inf A}^{\sup A} \frac{\widetilde{m}(\boldsymbol{n} - \boldsymbol{e}^A_{\boldsymbol{\xi}} + \boldsymbol{e}^{A_{\leq x}}_{\boldsymbol{\xi}} + \boldsymbol{e}^{A_{>x}}_{\boldsymbol{\xi}})}{\widetilde{m}(\boldsymbol{n})} \nu(x)\mathrm{d}x,$$

where $A_{\leq x} = A \cap [0,x]$ and $A_{>x} = A \cap (x,1]$. The resulting state is $\boldsymbol{n} - \boldsymbol{e}^A_{\boldsymbol{\xi}} + \boldsymbol{e}^{A_{\leq x}}_{\boldsymbol{\xi}} + \boldsymbol{e}^{A_{>x}}_{\boldsymbol{\xi}}$, with $x \in [\inf A, \sup A]$ chosen by the probability distribution proportional to $\frac{\widetilde{m}(\boldsymbol{n} - \boldsymbol{e}^A_{\boldsymbol{\xi}} + \boldsymbol{e}^{A_{\leq x}}_{\boldsymbol{\xi}} + \boldsymbol{e}^{A_{>x}}_{\boldsymbol{\xi}})}{\widetilde{m}(\boldsymbol{n})} \nu(x)\mathrm{d}x$.

In this description, $\widetilde{m}(\cdot)$ is the limit as $L \to \infty$ of (29), in a sense made more precise in Appendix B. Since $|\boldsymbol{\xi}|$ is finite, the jump distribution due to mutation has finite support. As is shown in Appendix B, it is further concentrated on transitions to states of the form $\widehat{\boldsymbol{n}} = \boldsymbol{n} - \boldsymbol{e}^A_{\boldsymbol{\xi}} + \boldsymbol{e}^A_{\boldsymbol{\xi}\setminus\{x\}}$ such that $x \notin \boldsymbol{\zeta}$ for any $\boldsymbol{\zeta}$ and B with $\widehat{n}^B_{\boldsymbol{\zeta}} > 0$ (i.e. if a mutation occurs at x then in the resulting configuration no haplotype carries the mutant allele at site x—the process obeys the infinitely-many-sites assumption).

6. The case of no mutation

As noted in the Introduction, it is possible to make further progress in the absence of mutation. Here we study in further detail the (reduced) L-locus model with $\theta = 0$. One must take care; the diffusion is no longer ergodic and the stationary distribution is not unique. In fact any distribution placing all its mass at $\delta_{\boldsymbol{j}}$ for some $\boldsymbol{j} \in E$ is an invariant distribution for $\boldsymbol{X}$; one haplotype $\boldsymbol{j}$ ultimately becomes fixed in the population, and once the diffusion hits this state it stays there. Nevertheless, for each invariant distribution we can find a non-trivial dual process. Here we adapt the results of Section 3.2. In order to normalise the duality function of (28) with respect to $\boldsymbol{X}_\infty \sim \delta_{\boldsymbol{j}}$, it is clear that $n^A_{\boldsymbol{i}}$ can be nonzero only if $\boldsymbol{i} = \boldsymbol{j}|_A$, and then (28) simplifies to

$$\widetilde{F}(\boldsymbol{x}, \boldsymbol{n}) = \prod_{\emptyset \neq A \subseteq [L]} (x^A_{\boldsymbol{j}})^{n^A}.$$

From this one immediately obtains the transition rates of the dual process:

Coalescence. For each nonempty $A, B \subseteq [L]$, the process jumps to $\boldsymbol{n} - \boldsymbol{e}_{\boldsymbol{j}}^{A} - \boldsymbol{e}_{\boldsymbol{j}}^{B} + \boldsymbol{e}_{\boldsymbol{j}}^{A\cup B}$ at rate

$$\frac{1}{2} n^A (n^B - \delta_{AB}).$$

Recombination. For each nonempty $A \subseteq [L]$ and $l = \min A, \ldots, \max A - 1$, the process jumps to $\boldsymbol{n} - \boldsymbol{e}_{\boldsymbol{j}}^{A} + \boldsymbol{e}_{\boldsymbol{j}}^{A_{\leq l}} + \boldsymbol{e}_{\boldsymbol{j}}^{A_{>l}}$ at rate

$$\frac{1}{2} n^A \rho_l.$$

The state space is $\{\boldsymbol{n} \in \Xi_{E,n} : n_{\boldsymbol{i}}^{A} = 0 \text{ if } \boldsymbol{i} \neq \boldsymbol{j}|_A\}$. This process describes the way that ancestral material is dispersed across the ancestors of a sample. It is the number of lineages in a (reduced, L-locus) ARG. For $L = 2$, the dynamics of this process are studied by, for example, Griffiths (1991) and Simonsen and Churchill (1997). Note that the degree of $\boldsymbol{n}$ is non-increasing, and, assuming that each locus is represented at least once in the initial sample, the process reaches a stationary state with support $\{\text{degree}(\boldsymbol{n}) = L\}$ (each locus has precisely one ancestor), with $\{n = 1\}$ a recurrent set (one individual is simultaneously ancestral at all loci). Starting from a single individual, ancestral material fragments back in time across many different individuals, before almost surely reconvening again within a single ancestor. Esser et al. (2016) call this the *partitioning process* in the context of the Moran model. In the same context, Bobrowski et al. (2010) study its rate of convergence to stationarity and provide a computer program to compute its transient distribution. Wiuf and Hein (1997) study the process in the context of the continuous model of Section 5, where they use it to address the question of how many genetic ancestors there are to a contemporary human chromosome.

It is convenient to denote the partitions directly. That is, if $\widetilde{\boldsymbol{L}}$ evolves as a partitioning process (with $\text{degree}(\widetilde{\boldsymbol{L}}_0) = L$), then let $\Theta_t = \{A \subseteq [L] : \widetilde{L}_t^A = 1\}$. Further writing

$$x_{\boldsymbol{j}}^{\Theta} = \prod_{A \in \Theta} x_{\boldsymbol{j}}^{A},$$

for a partition Θ, the duality equation can be written concisely as

$$\mathbb{E}\left[(X_{\boldsymbol{j}}^{\Phi})_t \mid \boldsymbol{X}_0 = \boldsymbol{x}\right] = \mathbb{E}\left[x_{\boldsymbol{j}}^{\Theta_t} \mid \Theta_0 = \Phi\right]. \tag{36}$$

It relates two particularly important quantities. Expectation on the left-hand side is with respect to $\boldsymbol{X}$ evolving forward in time according to (19) (with $\theta = 0$). The left-hand side is therefore a transient moment of the Wright-Fisher diffusion involving combinations of the alleles comprising the haplotype $\boldsymbol{j}$, where the combinations of interest are specified by a partition Φ. Expectation on the right-hand side is with respect to $\boldsymbol{\Theta} = (\Theta_t)_{t\geq 0}$ evolving backward in time from Φ. The right-hand side is therefore the PGF for the configuration of lineages in a reduced ARG. Mano (2013) uses the relationship between these quantities to find, among other things, the probability distribution of Θ_t for $L = 2$. Via a change of co-ordinate system, Esser et al. (2016) find the distribution of Θ_t for $L = 3$.

Letting $t \to \infty$ in (36) is also instructive. We find

$$\mathbb{E}\left[(X_{\boldsymbol{j}}^{\Phi})_{\infty} \mid \boldsymbol{X}_0 = \boldsymbol{x}\right] = \mathbb{E}\left[x_{\boldsymbol{j}}^{\Theta_\infty} \mid \Theta_0 = \Phi\right]. \tag{37}$$

The left-hand side of (37) is

$$\begin{aligned}\mathbb{E}\left[(X_{\boldsymbol{j}}^{\Phi})_{\infty} \mid \boldsymbol{X}_0 = \boldsymbol{x}\right] &= \mathbb{P}[(X_{\boldsymbol{j}}^{A})_{\infty} = 1, \forall A \in \Phi \mid \boldsymbol{X}_0 = \boldsymbol{x}] \\ &= \mathbb{P}[\boldsymbol{X}_\infty = \boldsymbol{e}_{\boldsymbol{j}} \mid \boldsymbol{X}_0 = \boldsymbol{x}],\end{aligned}$$

the probability that the haplotype $\boldsymbol{j}$ ultimately fixes in the population, starting from initial frequencies $\boldsymbol{x}$. The right-hand side of (37) is the PGF of Θ_∞, the stationary distribution of the partitioning process. Notice that both sides of (37) are independent of Φ. Notice also that, although the left-hand side is conditioned on the initial frequencies $\boldsymbol{x}$ of all haplotypes, it is only terms of the form $x_{\boldsymbol{j}}^{A}$ which are needed—the marginal frequency of haplotypes agreeing with $\boldsymbol{j}$ at a subset A of loci. Frequencies of alleles not appearing in $\boldsymbol{j}$ are immaterial (except through their aggregate frequency, which is expressible in terms of $x_{\boldsymbol{j}}^{A}$). Thus, for the purpose of computing (37), at each given locus l one could aggregate all alleles not equal to j_l and treat them as a single type with frequency $1 - x_{j_l}^{\{l\}}$.

The above reasoning motivates our interest in Θ_∞ in providing multilocus fixation probabilities. Let us spell this out further. First note that the fixation probability can be expressed as

$$\mathbb{P}[\boldsymbol{X}_\infty = \boldsymbol{e}_{\boldsymbol{j}} \mid \boldsymbol{X}_0 = \boldsymbol{x}] = \sum_{\Phi} \mathbb{F}(\Phi) x_{\boldsymbol{j}}^{\Phi},$$

where $\mathbb{F}(\Phi)$ is the probability that there are $|\Phi|$ single individuals whose descendents cause the haplotype $\boldsymbol{j}$ to fix according to the partition Φ; that

is, if ϕ_k is the kth block of Φ then the kth of the $|\Phi|$ individuals is the ancestor to the whole population at the loci in ϕ_k, and this individual has haplotype in agreement with $\boldsymbol{j}$ at these loci. Writing out both sides of (37),

$$\sum_{\Phi} \mathbb{F}(\Phi) x_{\boldsymbol{j}}^{\Phi} = \sum_{\Phi} \mathbb{P}(\Theta_{\infty} = \Phi) x_{\boldsymbol{j}}^{\Phi}, \tag{38}$$

and therefore

$$\mathbb{F}(\Phi) = \mathbb{P}(\Theta_{\infty} = \Phi). \tag{39}$$

We emphasise that (39) is a nice consequence of duality. In words, the stationary probability that the ancestors of the population partition the loci according to Φ is equal to the probability that $|\Phi|$ individuals fix according to the partition Φ. This argument could be extended to a continuous model of a gene as in Section 5, in which case Φ is a partition of $[0, 1]$.

Consider as a simple example the case of $L = 2$ loci. There are two possible partitions, $\{\{1, 2\}\}$ and $\{\{1\}, \{2\}\}$. Numbering these states as 1 and 2, the transition rate matrix of $\boldsymbol{\Theta}$ is

$$\widetilde{\boldsymbol{Q}} = \begin{pmatrix} -\rho_1/2 & \rho_1/2 \\ 1 & -1 \end{pmatrix}.$$

The distribution of Θ_{∞} is the unit solution $\boldsymbol{\pi}$ to $\boldsymbol{\pi}\widetilde{\boldsymbol{Q}} = \boldsymbol{0}$, which is easily verified to be

$$\boldsymbol{\pi} = \left(\frac{2}{2 + \rho_1}, \frac{\rho_1}{2 + \rho_1} \right).$$

The right-hand side of (37) is

$$\frac{2}{2 + \rho_1} x_{\boldsymbol{j}} + \frac{\rho_1}{2 + \rho_1} x_{\boldsymbol{j}}^{\{1\}} x_{\boldsymbol{j}}^{\{2\}},$$

and by duality this is the probability of fixation of $\boldsymbol{j}$ when initial frequencies are $\boldsymbol{x}$. If the population is initially at linkage equilibrium, so that $x_{\boldsymbol{j}} = x_{\boldsymbol{j}}^{\{1\}} x_{\boldsymbol{j}}^{\{2\}}$, then (36) becomes

$$\mathbb{E}[(X_{\boldsymbol{j}}^{\Phi})_t \mid \boldsymbol{X}_0 = \boldsymbol{x}] = x_{\boldsymbol{j}}^{\{1\}} x_{\boldsymbol{j}}^{\{2\}},$$

because $x_{\boldsymbol{j}}^{\Theta_t} = x_{\boldsymbol{j}}^{\{1\}} x_{\boldsymbol{j}}^{\{2\}}$ for all Θ_t. This agrees with our intuition that fixation probabilities of the two loci are independent when the initial state is one of

linkage equilibrium. Of course, a similar statement can be made for more than two loci.

The stationary distribution of Θ for $L = 3$ loci is given by Wiuf and Hein (1997), and its transient dynamics are studied by Esser et al. (2016), who also found an analogue of (39) for a two-locus Moran model.

6.1. The stationary distribution of the partitioning process

While the stationary distribution $\boldsymbol{\pi}$ of Θ_∞ is of interest, solving $\boldsymbol{\pi}\widetilde{\boldsymbol{Q}} = \mathbf{0}$ may not be straightforward because the size of this linear system grows rapidly with L. More precisely, the state space for Θ_t is the set of partitions of $[L]$. The number of such partitions is B_L, the Lth Bell number, which grows at least exponentially with L. In this subsection we show how one can compute the stationary distribution of Θ_∞ by solving a much smaller system, provided one has already computed the corresponding solution for an $(L-1)$-locus system. In this subsection we will use the superscript (L) to denote the dependence on L.

The key idea is to consider the collection of indicators $\boldsymbol{\epsilon}^{(L)} := (\epsilon_{ij})_{i,j\in[L]}$ defined by

$$\epsilon_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are in the same block of } \Theta_\infty, \\ 0 & \text{otherwise.} \end{cases}$$

Then $\boldsymbol{\pi}^{(L)}$ is expressible as a vector of joint moments of $\boldsymbol{\epsilon}^{(L)}$. For example, if $L = 2$ then $\boldsymbol{\pi}^{(2)} = \mathbb{E}(\epsilon_{12}, 1-\epsilon_{12})$. If $L = 3$ then

$$\begin{aligned} \boldsymbol{\pi}^{(3)\prime} &= \begin{pmatrix} \mathbb{P}(\Theta_\infty = \{\{1,2,3\}\}) \\ \mathbb{P}(\Theta_\infty = \{\{1,2\},\{3\}\}) \\ \mathbb{P}(\Theta_\infty = \{\{1,3\},\{2\}\}) \\ \mathbb{P}(\Theta_\infty = \{\{1\},\{2,3\}\}) \\ \mathbb{P}(\Theta_\infty = \{\{1\},\{2\},\{3\}\}) \end{pmatrix} = \mathbb{E}\begin{pmatrix} \epsilon_{12}\epsilon_{23} \\ \epsilon_{12}(1-\epsilon_{23}) \\ \epsilon_{13}(1-\epsilon_{12}) \\ (1-\epsilon_{12})\epsilon_{23} \\ (1-\epsilon_{12})(1-\epsilon_{13})(1-\epsilon_{23}) \end{pmatrix} \\ &= \mathbb{E}\begin{pmatrix} \epsilon_{12}\epsilon_{23} \\ \epsilon_{12} - \epsilon_{12}\epsilon_{23} \\ \epsilon_{13} - \epsilon_{12}\epsilon_{23} \\ \epsilon_{23} - \epsilon_{12}\epsilon_{23} \\ 1 - \epsilon_{12} - \epsilon_{13} - \epsilon_{23} + 2\epsilon_{12}\epsilon_{23} \end{pmatrix}. \end{aligned} \tag{40}$$

Some of the terms on the right-hand side of (40) are known from the two-locus solution:

$$\mathbb{E}[\epsilon_{12}] = \frac{2}{2+\rho_1}, \quad \mathbb{E}[\epsilon_{23}] = \frac{2}{2+\rho_2}, \quad \mathbb{E}[\epsilon_{13}] = \frac{2}{2+\rho_1+\rho_2}. \tag{41}$$

Substituting these results into $\boldsymbol{\pi}^{(3)}\widetilde{\boldsymbol{Q}}^{(3)} = \mathbf{0}$, the number of unknowns is reduced from $B_3 = 5$ down to just one, $\mathbb{E}[\epsilon_{12}\epsilon_{23}]$.

This idea extends to L loci. Suppose we have found $\boldsymbol{\pi}^{(L-1)}$; then we know all required joint moments of $\boldsymbol{\epsilon}^{(L-1)}$. The sequence $(\boldsymbol{\epsilon}^{(L)})_{L=1,2,\ldots}$ has an important consistency property: the marginal joint moments of $\boldsymbol{\epsilon}^{(L)}$ involving only the indices $1, 2, \ldots, L-1$ coincide with those of $\boldsymbol{\epsilon}^{(L-1)}$. Furthermore, by rescaling the recombination rate across any missing loci, we also know all the necessary joint moments of $\boldsymbol{\epsilon}^{(L)}$ involving indices with at most $L-1$ distinct entries in $1, 2, \ldots, L$. For example, by "forgetting" locus 2 we obtain $\mathbb{E}[\epsilon_{13}]$ in (41) by treating loci 1 and 3 as conforming to a two-locus model with recombination parameter $(\rho_1 + \rho_2)/2$. After exploiting this consistency property, the number of remaining unknown terms in $\boldsymbol{\pi}^{(L)}\widetilde{\boldsymbol{Q}}^{(L)} = \mathbf{0}$ is, we claim, equal to

$$S_L := (-1)^L + \sum_{k=1}^{L} (-1)^{k-1} B_{L-k}. \tag{42}$$

To see this, note that each unknown moment is of the form $\mathbb{E}[\epsilon_{i_1 j_1}\epsilon_{i_2 j_2}\cdots\epsilon_{i_d j_d}]$ in which each index $1, 2, \ldots, L$ appears at least once (otherwise we could appeal to the $(L-1)$-locus solution). Since each index is represented at least once, $\epsilon_{i_1 j_1}\epsilon_{i_2 j_2}\cdots\epsilon_{i_d j_d}$ defines a partition on $[L]$; that is, $\mathbb{E}[\epsilon_{i_1 j_1}\epsilon_{i_2 j_2}\cdots\epsilon_{i_d j_d}]$ corresponds uniquely to one entry in $\boldsymbol{\pi}^{(L)}$ (for example, when $L = 3$ we see from (40) that $\mathbb{E}[\epsilon_{12}\epsilon_{23}]$ is the first entry of $\boldsymbol{\pi}^{(3)}$). Moreover, this partition contains no singleton blocks, because any index i_k is paired in a block with some j_k. Thus, the number of unknown moments is equal to the number of partitions of $[L]$ containing no singleton blocks, which is given by (42) (A000296 of OEIS Foundation Inc., 2011, and references therein). By substituting known results from the $(L-1)$-locus solution for $\boldsymbol{\epsilon}^{(L-1)}$ into $\boldsymbol{\pi}^{(L)}\widetilde{\boldsymbol{Q}}^{(L)} = \mathbf{0}$ written in terms of moments of $\boldsymbol{\epsilon}^{(L)}$, the system is reduced from B_L to S_L equations, though S_L still exhibits exponential growth in L. The first few of these numbers are given in Table 1.

The above argument allows for the efficient computation of $\boldsymbol{\pi}^{(L)}$ successively for each L. The stationary distribution $\boldsymbol{\pi}^{(L)}$ is shown in Figure 2 for $L = 1, 2, \ldots, 6$, summarised by the stationary number of blocks $|\Theta_\infty|$ of Θ_∞. The complete solution for $\boldsymbol{\pi}^{(6)}$ is plotted in Figure 3 for a symmetric recombination model with $\rho_1 = \rho_2 = \cdots = \rho_5$. (Interestingly, the mode of $\boldsymbol{\pi}^{(6)}$ appears to be either $\{\{1, 2, 3, 4, 5, 6\}\}$ or $\{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}\}$ for any value of ρ_l.) We note that these observations are consistent with similar ones

Table 1: The number B_L of partitions of $[L]$, and the number S_L of partitions of $[L]$ containing no singleton blocks.

L	B_L	S_L
1	1	0
2	2	1
3	5	1
4	15	4
5	52	11
6	203	41
7	877	162
8	4140	715
9	21147	3425
10	115975	17722

made by Bobrowski et al. (2010, Section 4.1), who investigated Θ_∞ for a discrete-time Moran model by numerically iterating the partitioning process over generations until convergence to a chosen precision.

Duality tells us that fixation probabilities can be obtained as certain linear combinations of the curves in Figure 3. For example, suppose the population is fixed for a wild-type allele at each of the six loci. At each locus a mutant appears on the wild-type background and its haplotype drifts to frequency 1/6 (this might be thought of as a haplotype frequency configuration of maximal Hill-Robertson-type interference, though here everything is neutral). What is the probability that all six mutant alleles ultimately fix? Letting $\boldsymbol{j}$ denote the haplotype comprised of all six mutant alleles, from (38) the only partition Φ for which $x_{\boldsymbol{j}}^{\Phi}$ is nonzero is $\Phi = \{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\}\}$. Thus, (38) tells us that the fixation probability for $\boldsymbol{j}$ is given by the stationary probability of Φ (the dashed line in Figure 3) times $x_{\boldsymbol{j}}^{\Phi} = \left(\frac{1}{6}\right)^6$. So in this example, the dashed curve in Figure 3 also provides the fixation probability of $\boldsymbol{j}$ relative to the completely unlinked case, $\rho_l = \infty$.

7. Discussion

This paper makes three main contributions. First, we constructed the first duality relationships for population genetics models involving all of genetic

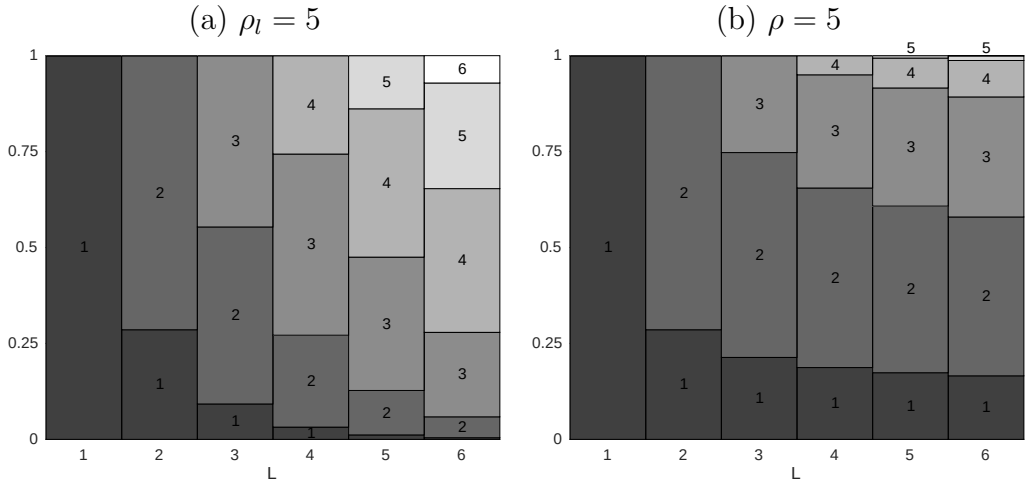
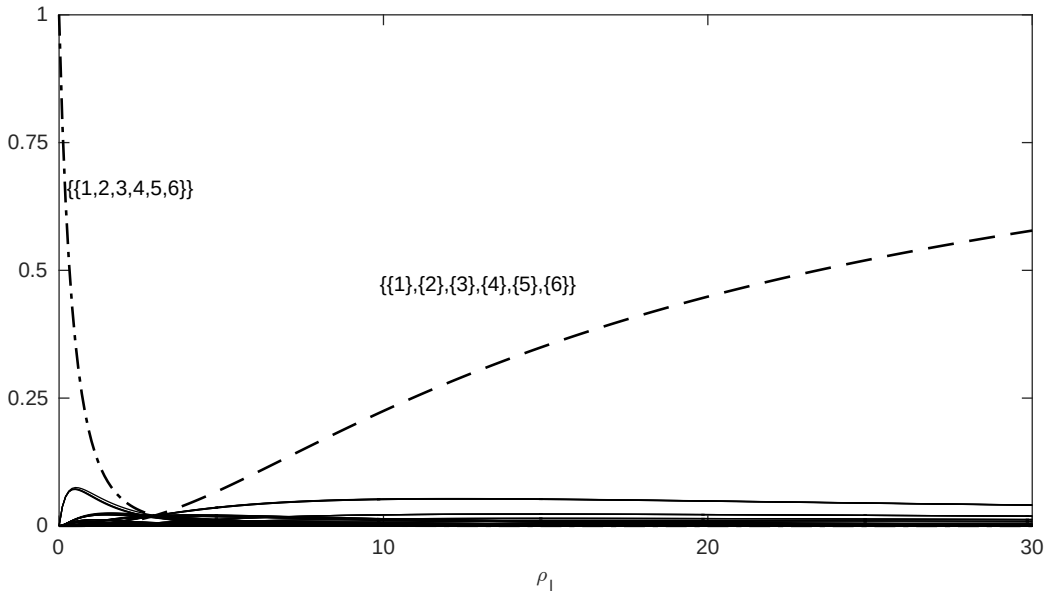
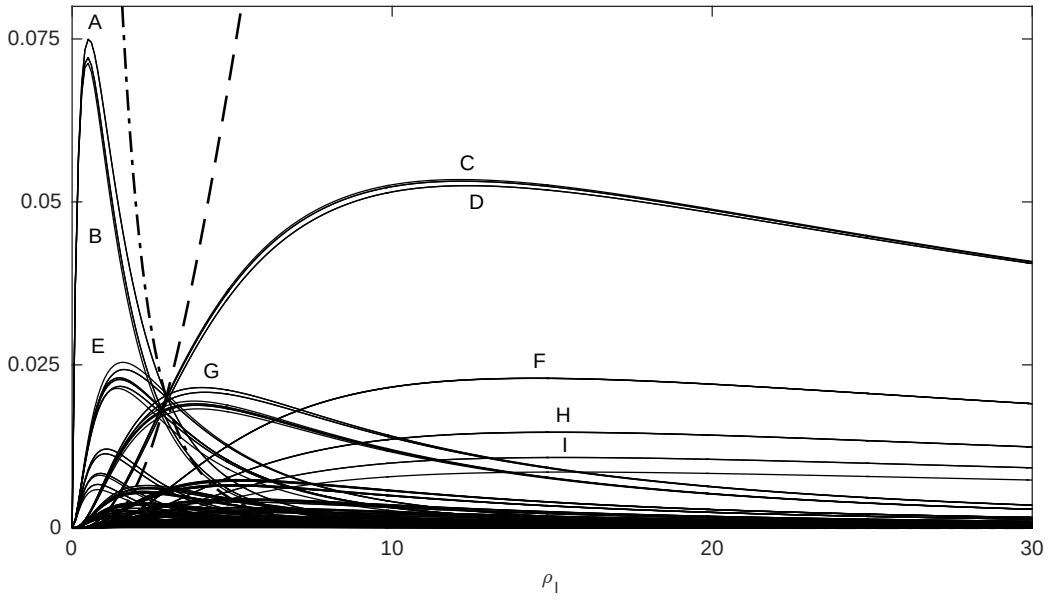


Figure 2: Stationary distribution of the number of fragments, $|\Theta_\infty|$, in an L-locus model. (a): Fixed per-locus recombination rate, $\rho_l = 5$. (b): Fixed total recombination rate, $\rho = \sum_{l=1}^{L-1} \rho_l = 5$.

drift, mutation, and recombination. They make precise the link between two individually well studied objects; namely, the Wright-Fisher diffusion with recombination and the ARG. This is done first for a discrete model of recombination and mutation and later on for a continuous limit model. Second, we emphasise the methods underlying our approach: it is particularly algebraically efficient to express the duality of two processes through their infinitesimal generators and to apply those generators to appropriate *generating functions*. Furthermore, this method is fairly automatic and avoids the pitfalls of the probabilistic arguments that are often invoked to address these types of questions. The price for this, one might argue, is that a biological interpretation of the results may be obscured. In this paper we have attempted to spell out how such biological interpretations can be recovered, by distilling mathematical expressions where possible to simple interpretable statements about conditional evolution. Third, we have highlighted the usefulness of our results via two applications: we obtained an expression for the transition function of the diffusion, and we showed how the partitioning process that arises when mutation is ignored can be related to predictions for haplotype fixation probabilities.


1
0.75
0.5
0.25
0
{{1,2,3,4,5,6}}
{{1},{2},{3},{4},{5},{6}}
0
10
20
30
ρ_{I}

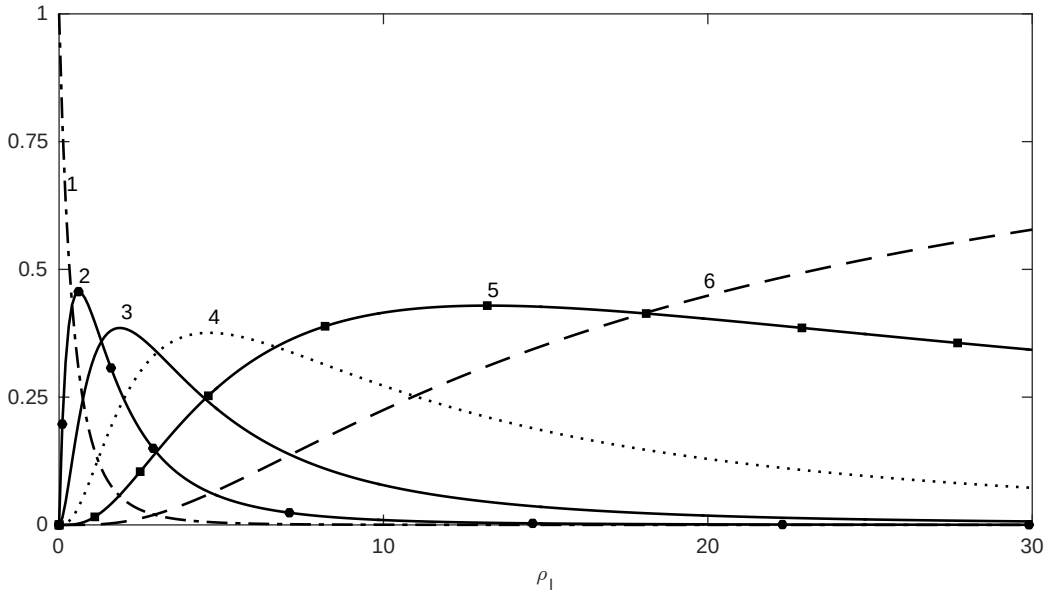
(a)


0.075
0.05
0.025
0
A
B
C
D
E
F
G
H
I
0
10
20
30
ρ_{I}

(b)

A	$\{\{1,2,3,4,5\},\{6\}\}$, $\{\{1\},\{2,3,4,5,6\}\}$.
B	$\{\{1,2,3,4\},\{5,6\}\}$, $\{\{1,2\},\{3,4,5,6\}\}$, $\{\{1,2,3\},\{4,5,6\}\}$.
C	$\{\{1\},\{2\},\{3,4\},\{5\},\{6\}\}$, $\{\{1\},\{2,3\},\{4\},\{5\},\{6\}\}$, $\{\{1\},\{2\},\{3\},\{4,5\},\{6\}\}$.
D	$\{\{1,2\},\{3\},\{4\},\{5\},\{6\}\}$, $\{\{1\},\{2\},\{3\},\{4\},\{5,6\}\}$.
E	$\{\{1\},\{2,3,4,5\},\{6\}\}$.
F	$\{\{1\},\{2\},\{3\},\{4,6\},\{5\}\}$, $\{\{1,3\},\{2\},\{4\},\{5\},\{6\}\}$, $\{\{1\},\{2,4\},\{3\},\{5\},\{6\}\}$, $\{\{1\},\{2\},\{3,5\},\{4\},\{6\}\}$.
G	$\{\{1\},\{2,3,4\},\{5\},\{6\}\}$, $\{\{1\},\{2\},\{3,4,5\},\{6\}\}$.
H	$\{\{1,4\},\{2\},\{3\},\{5\},\{6\}\}$, $\{\{1\},\{2\},\{3,6\},\{4\},\{5\}\}$, $\{\{1\},\{2,5\},\{3\},\{4\},\{6\}\}$.
I	$\{\{1,5\},\{2\},\{3\},\{4\},\{6\}\}$, $\{\{1\},\{2,6\},\{3\},\{4\},\{5\}\}$.

(c)



(d)

Figure 3: (a): Stationary fragment distribution, $\boldsymbol{\pi}^{(L)}$, of Θ_∞ for an $L = 6$ locus model with recombination parameter ρ_l at each breakpoint. (b): A detailed region of (a), with (c): a selection of partitions annotated. (d): The stationary distribution of the number of fragments, $|\Theta_\infty|$.

Acknowledgements

This work was supported in part by an Engineering & Physical Sciences Research Council grant to P.A.J. (EP/L018497/1). Part of this work was carried out while P.A.J. was at the University of California, Berkeley, supported in part by NIH Grant R01-GM094402, and while R.C.G. was visiting the Département de Mathématiques et de Statistique at the Université de Montréal, supported by the Clay Mathematics Institute. He would like to thank his hosts for their hospitality.

Appendix A. Useful identitites

For the function $\widetilde{S}(\boldsymbol{x}, \boldsymbol{n})$ defined by (23) and for $l = \min A, \ldots, \max A - 1$, note that

$$x_{\boldsymbol{i}}^{A\cup B}\widetilde{S}(\boldsymbol{x}, \boldsymbol{n} - \boldsymbol{e}_{\boldsymbol{i}}^{A} - \boldsymbol{e}_{\boldsymbol{i}}^{B}) = \frac{\binom{n-2}{\boldsymbol{n}-\boldsymbol{e}_{\boldsymbol{i}}^{A}-\boldsymbol{e}_{\boldsymbol{i}}^{B}}}{\binom{n-1}{\boldsymbol{n}-\boldsymbol{e}_{\boldsymbol{i}}^{A}-\boldsymbol{e}_{\boldsymbol{i}}^{B}+\boldsymbol{e}_{\boldsymbol{i}}^{A\cup B}}}\widetilde{S}(\boldsymbol{x}, \boldsymbol{n} - \boldsymbol{e}_{\boldsymbol{i}}^{A} - \boldsymbol{e}_{\boldsymbol{i}}^{B} + \boldsymbol{e}_{\boldsymbol{i}}^{A\cup B})$$
$$= \frac{n_{\boldsymbol{i}}^{A\cup B} + 1 - \delta_{A,A\cup B} - \delta_{B,A\cup B}}{n-1}\widetilde{S}(\boldsymbol{x}, \boldsymbol{n} - \boldsymbol{e}_{\boldsymbol{i}}^{A} - \boldsymbol{e}_{\boldsymbol{i}}^{B} + \boldsymbol{e}_{\boldsymbol{i}}^{A\cup B}), \quad \text{(A.1)}$$

$$x_{\boldsymbol{i}_{-l,j}}\widetilde{S}(\boldsymbol{x}, \boldsymbol{n} - \boldsymbol{e}_{\boldsymbol{i}}^{A}) = \frac{\binom{n-1}{\boldsymbol{n}-\boldsymbol{e}_{\boldsymbol{i}}^{A}}}{\binom{n}{\boldsymbol{n}-\boldsymbol{e}_{\boldsymbol{i}}^{A}+\boldsymbol{e}_{\boldsymbol{i}_{-l,j}}^{A}}}\widetilde{S}(\boldsymbol{x}, \boldsymbol{n} - \boldsymbol{e}_{\boldsymbol{i}}^{A} + \boldsymbol{e}_{\boldsymbol{i}_{-l,j}}^{A})$$
$$= \frac{n_{\boldsymbol{i}_{-l,j}}^{A} + 1 - \delta_{ilj}}{n}\widetilde{S}(\boldsymbol{x}, \boldsymbol{n} - \boldsymbol{e}_{\boldsymbol{i}}^{A} + \boldsymbol{e}_{\boldsymbol{i}_{-l,j}}^{A}), \quad \text{(A.2)}$$

$$x_{\boldsymbol{i}}^{A_{\le l}}x_{\boldsymbol{i}}^{A_{>l}}\widetilde{S}(\boldsymbol{x}, \boldsymbol{n} - \boldsymbol{e}_{\boldsymbol{i}}^{A}) = \frac{\binom{n-1}{\boldsymbol{n}-\boldsymbol{e}_{\boldsymbol{i}}^{A}}}{\binom{n+1}{\boldsymbol{n}-\boldsymbol{e}_{\boldsymbol{i}}^{A}+\boldsymbol{e}_{\boldsymbol{i}}^{A_{\le l}}+\boldsymbol{e}_{\boldsymbol{i}}^{A_{>l}}}}\widetilde{S}(\boldsymbol{x}, \boldsymbol{n} - \boldsymbol{e}_{\boldsymbol{i}}^{A} + \boldsymbol{e}_{\boldsymbol{i}}^{A_{\le l}} + \boldsymbol{e}_{\boldsymbol{i}}^{A_{>l}})$$
$$= \frac{(n_{\boldsymbol{i}}^{A_{\le l}} + 1)(n_{\boldsymbol{i}}^{A_{>l}} + 1)}{n(n+1)}\widetilde{S}(\boldsymbol{x}, \boldsymbol{n} - \boldsymbol{e}_{\boldsymbol{i}}^{A} + \boldsymbol{e}_{\boldsymbol{i}}^{A_{\le l}} + \boldsymbol{e}_{\boldsymbol{i}}^{A_{>l}}). \quad \text{(A.3)}$$

Appendix B. The continuous limit

In this appendix we show how to recover the continuous dual process described in Section 5 when the L-locus model is embedded in it; E_l, $\boldsymbol{P}^{(l)}$,

θ_l, and ρ_l are defined as in that section, and we let $L \to \infty$. To emphasise the dependence on L, in this appendix we will write $\boldsymbol{n}^{(L)}$, $\Xi_{E,n}^{(L)}$, and $\mathscr{L}^{(L)}$ for $\boldsymbol{n}$, $\Xi_{E,n}$, and $\mathscr{L}$. In order to identify the limiting behaviour of the process $\widetilde{\boldsymbol{L}}$ of Theorem 1, we proceed by fixing $\boldsymbol{n} \in \Xi_{[0,1],n}$, constructing a sequence $\boldsymbol{n}^{(L)} \in \Xi_{E,n}^{(L)}$ converging to $\boldsymbol{n}$ (in a manner to be defined precisely below), and then seeking the limit of $\mathscr{L}^{(L)}\widetilde{F}(\boldsymbol{x}, \boldsymbol{n}^{(L)})$ as $L \to \infty$.

To construct a sequence $(\boldsymbol{n}^{(L)})_{L\in\mathbb{N}}$ converging to some $\boldsymbol{n} \in \Xi_{[0,1],n}$, we define $\boldsymbol{n}^{(L)}$ as:

$$n_{\boldsymbol{\xi}^{(L)}}^{A^{(L)}} = \sum_{\substack{A\subseteq[0,1]:\\ A^{(L)}=LA\cap[L]}} \sum_{\substack{\boldsymbol{\xi}\subseteq A:\,|\boldsymbol{\xi}|<\infty,\\ \xi_i^{(L)}L=\lceil \xi_i L\rceil,\, i=0,1,\ldots}} n_{\boldsymbol{\xi}}^{A}. \tag{B.1}$$

Equation (B.1) defines an obvious 'coarsening' for representing a sample from the continuous model in its L-locus counterpart: the position of each mutant site is rounded up to the nearest multiple of $\frac{1}{L}$, and the segment A over which a haplotype is ancestral is represented by the collection $\{l \in [L] : \frac{l}{L} \in A\} =: A^{(L)}$. Given a sample $\boldsymbol{n}$, for sufficiently large L we have

$$n_{\boldsymbol{\xi}^{(L)}}^{A^{(L)}} = n_{\boldsymbol{\xi}}^{A}, \qquad \text{for each } \boldsymbol{\xi} \subseteq A \subseteq [0,1], \tag{B.2}$$

and we write $\boldsymbol{n}^{(L)} \to \boldsymbol{n}$ as $L \to \infty$. Similarly, we can fix the role of $\boldsymbol{x}$ by choosing $x_{\boldsymbol{\xi}^{(L)}}^{A^{(L)}} = x_{\boldsymbol{\xi}}^{A}$ for each $\boldsymbol{\xi}$ and A with $n_{\boldsymbol{\xi}}^{A} > 0$.

In this formulation, for sufficiently large L equation (27) becomes:

$$\begin{aligned}
\mathscr{L}^{(L)}\widetilde{S}(\boldsymbol{x}, \boldsymbol{n}^{(L)}) = {}& \\
& \frac{1}{2}\sum_{\emptyset\neq A\subseteq[0,1]}\Bigg[\sum_{\emptyset\neq B\subseteq[0,1]}\sum_{\boldsymbol{\xi}\subseteq(A\cup B)} n(n_{\boldsymbol{\xi}}^{A\cup B}+1-\delta_{A,A\cup B}-\delta_{B,A\cup B}) \\
& \qquad\qquad \times\widetilde{S}(\boldsymbol{x}, \boldsymbol{n}-\boldsymbol{e}_{\boldsymbol{\xi}}^{A}-\boldsymbol{e}_{\boldsymbol{\xi}}^{B}+\boldsymbol{e}_{\boldsymbol{\xi}}^{A\cup B}) \\
& + \theta\sum_{\boldsymbol{\xi}\subseteq A}\int_{\bigcup_{l\in A^{(L)}}[\frac{l-1}{L},\frac{l}{L}]} (n_{\tilde{\boldsymbol{\xi}}(x)}^{A}+1)\widetilde{S}(\boldsymbol{x}, \boldsymbol{n}-\boldsymbol{e}_{\boldsymbol{\xi}}^{A}+\boldsymbol{e}_{\tilde{\boldsymbol{\xi}}(x)}^{A})\eta(x)\mathrm{d}x \\
& + \rho\sum_{\boldsymbol{\xi}\subseteq A}\int_{\frac{1}{L}(\min A^{(L)}-1)}^{\frac{1}{L}\max A^{(L)}} \frac{(n_{\boldsymbol{\xi}}^{A_{\leq x}}+1)(n_{\boldsymbol{\xi}}^{A_{>x}}+1)}{n+1} \\
& \qquad\qquad \times\widetilde{S}(\boldsymbol{x}, \boldsymbol{n}-\boldsymbol{e}_{\boldsymbol{\xi}}^{A}+\boldsymbol{e}_{\boldsymbol{\xi}}^{A_{\leq x}}+\boldsymbol{e}_{\boldsymbol{\xi}}^{A_{>x}})\nu(x)\mathrm{d}x\Bigg]
\end{aligned}$$

$$- \left[n(n-1) + \sum_{A \subseteq [0,1]} n^A \left(\theta \int_{\bigcup_{l \in A^{(L)}} [\frac{l-1}{L}, \frac{l}{L}]} \eta(x) \mathrm{d}x \right.\right.$$
$$\left.\left. + \rho \int_{\frac{1}{L}(\min A^{(L)} - 1)}^{\frac{1}{L} \max A^{(L)}} \nu(x) \mathrm{d}x \right) \right] \widetilde{S}(\boldsymbol{x}, \boldsymbol{n}), \quad \text{(B.3)}$$

where $A_{\leq x} = A \cap [0, x]$, $A_{>x} = A \cap (x, 1]$, and $\bar{\boldsymbol{\xi}}(x)$ is given by (35). [Superscripts illustrating the dependence of $\boldsymbol{n}^{(L)}$ on L can be dropped, by virtue of (B.2).] We can now take the limit as $L \to \infty$ in (B.3); simply replace the range of integration for the mutation terms with A, and replace the range of integration for the recombination terms with $[\inf A, \sup A]$. In a similar manner, one can reformulate $\mathscr{L}^{(L)} \widetilde{F}(\boldsymbol{x}, \boldsymbol{n}^{(L)})$ and let $L \to \infty$ to find

$$\begin{aligned}
&\mathscr{L} \widetilde{F}(\boldsymbol{x}, \boldsymbol{n}) = \\
&\frac{1}{2} \sum_{\emptyset \neq A \subseteq [0,1]} \left[\sum_{\emptyset \neq B \subseteq [0,1]} \sum_{\boldsymbol{\xi} \subseteq (A \cup B)} n_{\boldsymbol{\xi}}^A (n_{\boldsymbol{\xi}}^B - \delta_{AB}) \frac{\widetilde{m}(\boldsymbol{n} - \boldsymbol{e}_{\boldsymbol{\xi}}^A - \boldsymbol{e}_{\boldsymbol{\xi}}^B + \boldsymbol{e}_{\boldsymbol{\xi}}^{A \cup B})}{\widetilde{m}(\boldsymbol{n})} \right. \\
&\qquad \left. \times \widetilde{F}(\boldsymbol{x}, \boldsymbol{n} - \boldsymbol{e}_{\boldsymbol{\xi}}^A - \boldsymbol{e}_{\boldsymbol{\xi}}^B + \boldsymbol{e}_{\boldsymbol{\xi}}^{A \cup B}) \right. \\
&+ \theta \sum_{\boldsymbol{\xi} \subseteq A} n_{\boldsymbol{\xi}}^A \int_A \frac{\widetilde{m}(\boldsymbol{n} - \boldsymbol{e}_{\boldsymbol{\xi}}^A + \boldsymbol{e}_{\bar{\boldsymbol{\xi}}(x)}^A)}{\widetilde{m}(\boldsymbol{n})} \widetilde{F}(\boldsymbol{x}, \boldsymbol{n} - \boldsymbol{e}_{\boldsymbol{\xi}}^A + \boldsymbol{e}_{\bar{\boldsymbol{\xi}}(x)}^A) \eta(x) \mathrm{d}x \\
&+ \rho \sum_{\boldsymbol{\xi} \subseteq A} n_{\boldsymbol{\xi}}^A \int_{\inf A}^{\sup A} \frac{\widetilde{m}(\boldsymbol{n} - \boldsymbol{e}_{\boldsymbol{\xi}}^A + \boldsymbol{e}_{\boldsymbol{\xi}}^{A_{\leq x}} + \boldsymbol{e}_{\boldsymbol{\xi}}^{A_{>x}})}{\widetilde{m}(\boldsymbol{n})} \right. \\
&\qquad \left. \left. \times \widetilde{F}(\boldsymbol{x}, \boldsymbol{n} - \boldsymbol{e}_{\boldsymbol{\xi}}^A + \boldsymbol{e}_{\boldsymbol{\xi}}^{A_{\leq x}} + \boldsymbol{e}_{\boldsymbol{\xi}}^{A_{>x}}) \nu(x) \mathrm{d}x \right] \right. \\
&- \left[n(n-1) + \sum_{A \subseteq [0,1]} n^A \left(\theta \int_A \eta(x) \mathrm{d}x + \rho \int_{\inf A}^{\sup A} \nu(x) \mathrm{d}x \right) \right] \widetilde{F}(\boldsymbol{x}, \boldsymbol{n}), \quad \text{(B.4)}
\end{aligned}$$

where $\widetilde{m}(\widehat{\boldsymbol{n}})/\widetilde{m}(\boldsymbol{n})$ is defined as the weak limit satisfying

$$\int_C \frac{\widetilde{m}(\widehat{\boldsymbol{n}})}{\widetilde{m}(\boldsymbol{n})} \widetilde{F}(\boldsymbol{x}, \widehat{\boldsymbol{n}}) \lambda(x) \mathrm{d}x = \lim_{L \to \infty} \int_C \frac{\widetilde{m}(\widehat{\boldsymbol{n}}^{(L)})}{\widetilde{m}(\boldsymbol{n}^{(L)})} \widetilde{F}(\boldsymbol{x}, \widehat{\boldsymbol{n}}^{(L)}) \lambda(x) \mathrm{d}x,$$

for $C \subseteq [0, 1]$ and λ a probability density on $[0, 1]$. (We refrain from passing the limit through the integral, since in some instances it is necessary to interpret the limit in a Dirac sense; see below.) The interpretation of (B.4) as

the generator of a pure jump Markov process is clear, and the terms corresponding to coalescence and recombination events agree with the description given in Section 5. The mutation term, however, reads as:

Mutation. For each $A \subseteq [0,1]$ and $\boldsymbol{\xi} \subseteq A$, the process jumps at rate

$$\frac{\theta}{2} n^A_{\boldsymbol{\xi}} \int_A \frac{\widetilde{m}(\boldsymbol{n} - \boldsymbol{e}^A_{\boldsymbol{\xi}} + \boldsymbol{e}^A_{\boldsymbol{\xi}(x)})}{\widetilde{m}(\boldsymbol{n})} \eta(x) \mathrm{d}x.$$

The resulting state is $\boldsymbol{n} - \boldsymbol{e}^A_{\boldsymbol{\xi}} + \boldsymbol{e}^A_{\boldsymbol{\xi}(x)}$, with the position $x \in A$ chosen by the probability distribution proportional to $\frac{\widetilde{m}(\boldsymbol{n} - \boldsymbol{e}^A_{\boldsymbol{\xi}} + \boldsymbol{e}^A_{\boldsymbol{\xi}(x)})}{\widetilde{m}(\boldsymbol{n})} \eta(x) \mathrm{d}x$.

It remains to reconcile this with the description for mutation given in Section 5, which follows if we can show that the infinitely-many-sites assumption holds in the limit. More precisely, we should see transitions only to states of the form $\widehat{\boldsymbol{n}} = \boldsymbol{n} - \boldsymbol{e}^A_{\boldsymbol{\xi}} + \boldsymbol{e}^A_{\boldsymbol{\xi} \setminus \{x\}}$ such that $x \in \boldsymbol{\xi}$, and such that $x \notin \boldsymbol{\zeta}$ for any $\boldsymbol{\zeta}$ and B with $\widehat{n}^B_{\boldsymbol{\zeta}} > 0$. This holds by the following lemma, from which we deduce that if $\widehat{\boldsymbol{n}}$ is *not* of this form then $\widetilde{m}(\widehat{\boldsymbol{n}}^{(L)})/\widetilde{m}(\boldsymbol{n}^{(L)}) \to 0$ as $L \to \infty$.

Lemma 1. *Let*

$$s(\boldsymbol{n}^{(L)}) = \left| \bigcup_{\boldsymbol{\xi}^{(L)} \subseteq A^{(L)} : n^{A^{(L)}}_{\boldsymbol{\xi}^{(L)}} > 0} \boldsymbol{\xi}^{(L)} \right|$$

denote the total number of segregating sites in a sample $\boldsymbol{n}^{(L)} \in \Xi^{(L)}_{E,n}$. *If* $s(\boldsymbol{n}^{(L)}) = O(1)$ *then* $\widetilde{m}(\boldsymbol{n}^{(L)}) = O(L^{-s(\boldsymbol{n}^{(L)})})$ *as* $L \to \infty$.

Proof. $\widetilde{m}(\boldsymbol{n}^{(L)})$ satisfies the finite system (30), whose solution is unique. (The boundary condition is adjusted to account for our definition of $\boldsymbol{\xi}$ with respect to a reference haplotype: $\widetilde{m}(\boldsymbol{e}_{\boldsymbol{\xi}^{(L)}}) = \delta_{\boldsymbol{\xi}^{(L)}\emptyset}$.) It is straightforward to check that $\widetilde{m}(\boldsymbol{n}^{(L)}) = O(L^{-s(\boldsymbol{n}^{(L)})})$ satisfies this system: The left-hand side, and the first and third terms on the right are all clearly $O(L^{-s(\boldsymbol{n}^{(L)})})$. The second term on the right, corresponding to mutation events, has three contributions: First, there are $O(1)$ summands for which (in the notation of this section) $\widehat{\boldsymbol{n}} = \boldsymbol{n} - \boldsymbol{e}^{A^{(L)}}_{\boldsymbol{\xi}^{(L)}} + \boldsymbol{e}^{A^{(L)}}_{\tilde{\boldsymbol{\xi}}^{(L)}(\frac{l}{L})}$ has one fewer segregating site; these terms contribute $\theta_l \times \widetilde{m}(\widehat{\boldsymbol{n}}) = O(L^{-1} \times L^{-(s(\boldsymbol{n}^{(L)})-1)}) = O(L^{-s(\boldsymbol{n}^{(L)})})$. Second, there are $O(1)$ summands for which $\widehat{\boldsymbol{n}}$ has the same number of segregating sites (parallel mutations); these terms contribute $\theta_l \times \widetilde{m}(\widehat{\boldsymbol{n}}) = O(L^{-1} \times L^{-s(\boldsymbol{n}^{(L)})}) =$

$O(L^{-(s(\boldsymbol{n}^{(L)})+1)})$ and vanish in the limit. Third, there are $O(L)$ summands for which $\widehat{\boldsymbol{n}}$ has one extra segregating site (back mutations); these terms each contribute $O(L^{-1} \times L^{-s(\boldsymbol{n}^{(L)})+1})$ and also vanish in the limit. □

Thus, back mutations are not seen in the limit because the integrand in (34) vanishes, while parallel mutation are not seen because the integrand is $O(1)$ but the range of integration for such events has Lebesgue measure zero. The integral (34) is recognised retrospectively as a sum over at most $|\boldsymbol{\xi}|$ atoms.

References

Barbour, A. D., Ethier, S. N., Griffiths, R. C., 2000. A transition function expansion for a diffusion model with selection. Annals of Applied Probability 10 (1), 123–162.

Bobrowski, A., Wojdyła, T., Kimmel, M., 2010. Asymptotic behavior of a Moran model with mutations, drift and recombination among multiple loci. Journal of Mathematical Biology 61, 455–473.

Donnelly, P., Kurtz, T. G., 1999. Genealogical processes for Fleming-Viot models with selection and recombination. Annals of Applied Probability 9 (4), 1091–1148.

Donnelly, P., Tavaré, S., 1987. The population genealogy of the infinitely-many neutral alleles model. Journal of Mathematical Biology 25, 381–391.

Esser, M., Probst, S., Baake, E., 2016. Partitioning, duality, and linkage disequilibria in the Moran model with recombination. Journal of Mathematical Biology 73 (1), 161–197.

Etheridge, A. M., Griffiths, R. C., 2009. A coalescent dual process in a Moran model with genic selection. Theoretical Population Biology 75, 320–330.

Etheridge, A. M., Griffiths, R. C., Taylor, J. E., 2010. A coalescent dual process in a Moran model with genic selection, and the lambda coalescent limit. Theoretical Population Biology 78, 77–92.

Ethier, S. N., Griffiths, R. C., 1987. The infinitely-many-sites model as a measure-valued diffusion. The Annals of Probability 15 (2), 515–545.

Ethier, S. N., Griffiths, R. C., 1990a. The neutral two-locus model as a measure-valued diffusion. Advances in Applied Probability 22 (4), 773–786.

Ethier, S. N., Griffiths, R. C., 1990b. On the two-locus sampling distribution. Journal of Mathematical Biology 29, 131–159.

Ethier, S. N., Griffiths, R. C., 1993. The transition function of a Fleming-Viot process. Annals of Probability 21 (3), 1571–1590.

Ethier, S. N., Kurtz, T. G., 1993. Fleming-Viot processes in population genetics. SIAM Journal of Control and Optimization 31 (2), 345–386.

Fearnhead, P., 2002. The common ancestor at a nonneutral locus. Journal of Applied Probability 39, 38–54.

Fearnhead, P., 2003. Haplotypes: the joint distribution of alleles at linked loci. Journal of Applied Probability 40, 505–512.

Fearnhead, P., Donnelly, P., 2001. Estimating recombination rates from population genetic data. Genetics 159, 1299–1318.

Golding, G. B., 1984. The sampling distribution of linkage disequilibrium. Genetics 108, 257–274.

Griffiths, R. C., 1979. A transition density expansion for a multi-allele diffusion model. Advances in Applied Probability 11 (2), 310–325.

Griffiths, R. C., 1980. Lines of descent in the diffusion approximation of neutral Wright-Fisher models. Theoretical Population Biology 17, 37–50.

Griffiths, R. C., 1981. Neutral two-locus multiple allele models with recombination. Theoretical Population Biology 19, 169–186.

Griffiths, R. C., 1991. The two-locus ancestral graph. In: Basawa, I. V., Taylor, R. L. (Eds.), Selected proceedings of the Sheffield symposium on applied probability: 18. IMS Lecture Notes—Monograph series. Vol. 18. pp. 100–117.

Griffiths, R. C., Jenkins, P. A., Song, Y. S., 2008. Importance sampling and the two-locus model with subdivided population structure. Advances in Applied Probability 40 (2), 473–500.

Griffiths, R. C., Marjoram, P., 1996. Ancestral inference from samples of DNA sequences with recombination. Journal of Computational Biology 3 (4), 479–502.

Griffiths, R. C., Marjoram, P., 1997. An ancestral recombination graph. In: Donnelly, P., Tavaré, S. (Eds.), Progress in population genetics and human evolution. Vol. 87. Springer-Verlag Berlin, pp. 257–270.

Handa, K., 2002. Quasi-invariance and reversibility in the Fleming-Viot process. Probability Theory and Related Fields 122, 545–566.

Hudson, R. R., 1983. Properties of a neutral allele model with intragenic recombination. Theoretical Population Biology 23, 183–201.

Jansen, S., Kurt, N., 2014. On the notion(s) of duality for Markov processes. Probability Surveys 11, 59–120.

Jenkins, P. A., Griffiths, R. C., 2011. Inference from samples of DNA sequences using a two-locus model. Journal of Computational Biology 18 (1), 109–127.

Jenkins, P. A., Song, Y. S., 2009. Closed-form two-locus sampling distributions: accuracy and universality. Genetics 183, 1087–1103.

Kamm, J. A., Spence, J. P., Chan, J., Song, Y. S., 2016. Two-locus likelihoods under variable population size and fine-scale recombination rate estimation. Genetics 203 (3), 1381–1399.

Kingman, J. F. C., 1982. The coalescent. Stochastic Processes and their Applications 13 (3), 235–248.

Krone, S. M., Neuhauser, C., 1997. Ancestral processes with selection. Theoretical Population Biology 51 (3), 210–237.

Larribe, F., Lessard, S., 2008. A composite-conditional-likelihood approach for gene mapping based on linkage disequilibrium in windows of marker loci. Statistical Applications in Genetics and Molecular Biology 7 (1), Article 27.

Larribe, F., Lessard, S., Schork, N. J., 2002. Gene mapping via the ancestral recombination graph. Theoretical Population Biology 62, 215–229.

Lohse, K., Chmelik, M., Martin, S. H., Barton, N. H., 2016. Efficient strategies for calculating blockwise likelihoods under the coalescent. Genetics 202 (2), 775–786.

Lohse, K., Harrison, R. J., Barton, N. H., 2011. A general method for calculating likelihoods under the coalescent process. Genetics 189, 977–987.

Mano, S., 2013. Duality between the two-locus Wright-Fisher diffusion model and the ancestral process with recombination. Journal of Applied Probability 50, 256–271.

Neuhauser, C., Krone, S. M., 1997. The genealogy of samples in models with selection. Genetics 145, 519–534.

OEIS Foundation Inc., 2011. The on-line encyclopedia of integer sequences. URL `http://oeis.org`

Simonsen, K. L., Churchill, G. A., 1997. A Markov chain model of coalescence with recombination. Theoretical Population Biology 52, 43–59.

Stephens, M., 2007. Inference under the coalescent. In: Balding, D., Bishop, M., Cannings, C. (Eds.), Handbook of Statistical Genetics. Wiley, Chichester, UK, Ch. 26, pp. 878–908.

Stephens, M., Donnelly, P., 2003. Ancestral inference in population genetics models with selection. Australia and New Zealand Journal of Statistics 45 (3), 395–430.

Wiuf, C., Hein, J., 1997. On the number of ancestors to a DNA sequence. Genetics 147, 1459–1468.

Wright, S., 1949. Adaptation and selection. In: Jepson, G. L., Mayr, E., Simpson, G. G. (Eds.), Genetics, Paleontology and Evolution. Princeton University Press, Princeton, pp. 365–389.